GEOLOGICAL SURVEY OF FINLAND
Report of Investigation 217
2015

New computational methods for efficient utilisation of public data

Ala-Ilomäki, J., Cohen, J., Heilimo, J., Hyvönen, E., Hänninen, P., Ikonen, J., Middleton, M., Nevalainen, P., Pahikkala, T., Pohjankukka, J., Pulliainen, J., Riihimäki, H., Sutinen, R., Tuominen, S. and Varjo, J.

GTK

GEOLOGIAN TUTKIMUSKESKUS **GEOLOGICAL SURVEY OF FINLAND**

Tutkimusraportti 217 Report of Investigation 217

Ala-Ilomäki, J., Cohen, J., Heilimo, J., Hyvönen, E., Hänninen, P., Ikonen, J., Middleton, M., Nevalainen, P., Pahikkala, T., Pohjankukka, J., Pulliainen, J., Riihimäki, H., Sutinen, R., Tuominen, S. and Varjo, J.

NEW COMPUTATIONAL METHODS FOR EFFICIENT UTILISATION OF PUBLIC DATA

Front cover: Reference measurements for terrain trafficability application at the Pieksämäki harvesting site.
Photo: Matti Sirén, LuKe.

Unless otherwise indicated, the figures have been prepared by the authors of the publication.

ISBN 978-952-217-332-4 (PDF)
ISSN 0781-4240

Layout: Elvi Turtiainen Oy

Espoo 2015

Ala-Ilomäki, J.[1], Cohen, J.[2], Heilimo, J.[2], Hyvönen, E.[3], Hänninen, P.[4], Ikonen, J.[2], Middleton, M.[3], Nevalainen, P.[5], Pahikkala, T.[5], Pohjankukka, J.[5], Pulliainen, J.[2], Riihimäki, H.[1], Sutinen, R.[3], Tuominen, S.[1] & Varjo, J.[6] 2015. New computational methods for efficient utilisation of public data. *Geological Survey of Finland. Report of Investigation 217*, 55 pages, 32 figures, 5 tables and 1 appendices.

The project investigated the possibilities of publicly available spatial data in mapping and predicting of geospatial phenomena with economical importance. The aim was to develop practical applications based merely on open data from the databases of Finnish Meteorological Institute (FMI), Geological Survey of Finland (GTK), Finnish Forest Research Institute (Metla; since 1st Jan., 2015, Natural Resources Institute Finland, Luke) and National Land Survey of Finland (NLS).

Geographic Information Systems (GIS), Remote Sensing (RS) and Machine Learning (ML) techniques were applied in developing various applications. The most promising applications were: i) Hydrological Operations and Prediction Model, HOPS, ii) Mapping of mass-flow aggregate deposits for infrastructure construction, iii) Quick response mapping of forest floods, and, iv) Mapping of drainage networks. The HOPS is already an operational application, while other applications still need further validation to become operational. The results obtained from the other three potential applications were nevertheless encouraging.

The detection of storm damage using X-band SAR satellite data was not successful. The analysis may be improved by time-series imagery and multitude of frequency bands both of which are publicly available via Sentinel satellite series starting Dec. 2014. Predicting trafficability in forests needs further research including new physical models and high quality reference data regarding load bearing capacity of soils. However, the prediction accuracy of the route selection was good. In the future trafficability analysis should be developed by gathering the learning data online from forest harvester CAN-bus.

Future research should also focus on large scale texture analysis methods on spatial information and in the development of physical prediction models, regarding the utilisation of publicly open data. For example, the HOPS model can be implemented as a dynamical input component in future trafficability applications. The value of airborne laser scanning (ALS) is evident since it was an essential data source in many of the potential applications. Therefore, we stress the need of frequently updated ALS data, preferably with higher pulse density than currently, because it would benefit many applications via enhanced signal of the targets, e.g. ditches and boulders, and may open completely new possibilities for further development of applications.

Keywords (GeoRef Thesaurus, AGI): data management, data integration, geographic information systems, geophysical methods, airborne methods, laser scanning, radar methods, pattern recognition, Finland

[1] *Natural Resources Institute Finland (Luke), Jokiniemenkuja 1, FI-01370 Vantaa, Finland*
E-mail: jari.ala-ilomaki@luke.fi, henri.riihimaki@luke.fi, sakari.tuominen@luke.fi
[2] *Finnish Meteorological Institute, P.O. Box 503, FI-00101, Helsinki, Finland*
E-mail: juval.cohen@fmi.fi, jyri.heilimo@fmi.fi, jaakko.ikonen@fmi, jouni.pulliainen@fmi.fi
[3] *Geological Survey of Finland, P.O. Box 77, FI-96101 Rovaniemi, Finland*
E-mail: eija.hyvonen@gtk.fi, maarit.middleton@gtk.fi, raimo.sutinen@gtk.fi
[4] *Geological Survey of Finland, P.O. Box 96, FI-02151 ESPOO, Finland.*
E-mail: pekka.hanninen@gtk.fi
[5] *Department of Information Technology, FI-20014 Turku University, Finland*
E-mail: paavo.t.nevalainen@utu.fi, aatapa@utu.fi, jjepoh@utu.fi
[6] *Natural Resources Institute Finland (Luke), Viikinkaari 4, FI-00790 Helsinki, Finland*
E-mail: jari.varjo@luke.fi

Ala-Ilomäki, J.[1], Cohen, J.[2], Heilimo, J.[2], Hyvönen, E.[3], Hänninen, P.[4], Ikonen, J.[2], Middleton, M.[3], Nevalainen, P.[5], Pahikkala, T.[5], Pohjankukka, J.[5], Pulliainen, J.[2], Riihimäki, H.[1], Sutinen, R.[3], Tuominen, S.[1] & Varjo, J.[6] 2015. New computational methods for efficient utilisation of public data. *Geologian tutkimuskeskus, Tutkimusraportti 217,* 55 sivua, 32 kuvaa, 5 taulukkoa ja 1 liitettä.

Tässä projektissa tutkittiin julkisesti saatavilla olevan spatiaalisen datan soveltuvuus kartoitukseen ja spatiaaliseen ennustamiseen ja niin, että sillä olisi taloudellista merkitystä. Tavoitteena oli kehittää käytännön sovelluksia niiden avoimien datojen perusteella, joita Ilmatieteen laitos, Geologian tutkimuskeskus, Metsäntutkimuslaitos (1.1.2015 lähtien Luonnonvarakeskus, LuKe) ja Maanmittauslaitos ovat tuottaneet.

Paikkatietojärjestelmää (GIS), kaukokartoitusta ja koneoppimista käytettiin uusien sovellusten kehittämiseen. Lupaavimmat sovellukset olivat seuraavat: i) HOPS (Hydrological Operations and Prediction Model), ii) massamuodostumien kartoitus rakentamisen tarpeisiin, iii) nopea metsien tulvakartoitus ja iv) kuivatusverkostojen kartoitus. HOPS on jo operatiivisessa käytössä, kun taas muiden sovellusten vahvistaminen operatiiviseen toimintaan on paikallaan. Nämä kolme sovellusta osoittautuivat joka tapauksessa merkityksellisiksi tulevaisuutta ajatellen.

X-kanavainen SAR-satelliittiaineisto ei osoittautunut onnistuneeksi metsätuhojen paikantamiseen. Analyysiä on mahdollista parantaa käyttämällä aikasarjoja ja monikanavaista tutka-aineistoa, joista molemmat tulivat julkisesti saataviksi Sentinel-satelliittiohjelman laukaisun myötä joulukuussa 2014. Metsien kulkukelpoisuuden ennustaminen edellyttää maaperän uusia fysikaalisia malleja ja korkealuokkaista maaperän kantavuuden referenssiaineistoa. Kulkukelpoisuuden reittiennuste osoittautui jo nyt kuitenkin hyväksi. Tulevaisuudessa kulkukelpoisuusanalyysin tulee pohjautua opetusaineistoon, jonka toiminnassa oleva metsäkone kokoaa.

Tulevaisuudessa tutkimuksen tulee keskittyä laajamittaiseen spatiaalisen tiedon tekstuurianalyysiin ja fysikaalisten ennustemallien kehittämiseen niin, että niissä hyödynnetään avointa dataa. Esimerkiksi HOPS-malli voidaan implementoida dynaamisena syöttökomponenttina kulkukelpoisuuden sovelluksissa tulevaisuudessa. Laserkeilauksen edut ovat kiistattomat, sillä keilausdata oli monessa sovelluksessa oleellinen tekijä. Kokemuksiemme perusteella keilauslentoja tulee jatkossa uusia ja pulssitiheyttä tulee nostaa, jotta niillä saadaan suuri hyöty esimerkiksi ojien sekä kivisyyden ja lohkareiden kartoittamiseen. Nämä todennäköisesti avaavat myös kokonaan uusien sovellusten kehittämisen tulevaisuudessa.

Asiasanat (Geosanasto, GTK): tiedonhallinta, tietojen yhdistäminen, paikkatietojärjestelmät, geofysikaaliset menetelmät, lentomittaukset, laserkeilaus, tutkaluotausmenetelmät, hahmontunnistus, Suomi

[1] Luonnonvarakeskus, Jokiniemenkuja 1, 01370 Vantaa,
Sähköposti: jari.ala-ilomaki@luke.fi, henri.riihimaki@luke.fi, sakari.tuominen@luke.fi
[2] Ilmatieteenlaitos, PL 503, 00101 Helsinki
Sähköposti: juval.cohen@fmi.fi, jyri.heilimo@fmi.fi, jaakko.ikonen@fmi, jouni.pulliainen@fmi.fi
[3] Geologian tutkimuskeskus, PL 77, 96101 Rovaniemi
Sähköposti: eija.hyvonen@gtk.fi, maarit.middleton@gtk.fi, raimo.sutinen@gtk.fi
[4] Geologian tutkimuskeskus, PL 96, 02151 ESPOO
Sähköposti: pekka.hanninen@gtk.fi
[5] Informaatioteknologian laitos, 20014 Turun yliopisto
Sähköposti: paavo.t.nevalainen@utu.fi, aatapa@utu.fi, jjepoh@utu.fi
[6] Luonnonvarakeskus, Viikinkaari 4, 00790 Helsinki
Sähköposti: jari.varjo@luke.fi

CONTENTS

TERMINOLOGY AND ABBREVIATIONS 5

1 OVERVIEW OF THE PROJECT 6
1.1 Project participants and their data resources 6
1.2 Goals 7
1.3 Project timetable and chronological outline of project activities 8
1.4 Executive Summary 10

2 DATA 13
2.1 Public open data 13
2.2 The forest inventory (MS-NFI, Metla) 14
2.3 Airborne geophysical data and Quaternary geological mapping data 15
2.4 Airborne Laser Scanning data and NLS topographic database 15
2.5 COSMO-SkyMed SAR data 16

3 PREDICTION PERFORMANCE 16
3.1 Data matrix and prediction 17
3.2 Computational tasks 19
3.3 Machine Learning methods 20

4 CASE STUDIES AND RESULTS 20
4.1 Terrain and forest road trafficability 20
4.2 Flood detection 31
4.3 Forest drainage network extraction based on airborne laser scanning 32
4.4 Hydrological Operations and Prediction System, HOPS 37
4.5 Storm damage 38
4.6 Recognition of mass-flow deposits for infrastructure construction 38

5 TECHNOLOGICAL POTENTIAL 42
5.1 Implementation 42
5.2 Computational tasks 44
5.3 Review of GIS workflow 45
5.4 Cloud environments 45

6 DISCUSSION 47
6.1 Deliverables 47
6.2 Summary 47
6.3 Future research 48
6.4 Future commercial applications 50

ACKNOWLEDGEMENTS 53

REFERENCES 53

TERMINOLOGY AND ABBREVIATIONS

AEM: Airbone electromagnetic: airborne electromagnetic data
AGR, Airborne gamma radiation: airborne gamma radiation data from potassium decay
ALS: Airborne Laser Scanning, a general term for many mapping techniques. See also LiDAR.
CAN-bus: Controller Area Network harvester
CI: Concordance Index, measures how monotonically faithfull the prediction, also C -index is in categorical choices.
DEM: Digital Elevation Model, ground approximation from LiDAR
Derived feature: computed from a primary feature in purpose of improving the prediction
Environment: feature vector cluster describing a category for the surrounding small area around a location. The term is close to its common language meaning.
EUREF-FIN: Co-ordinate system used in this project.
Feature: numerical or categorical value associated to a map position. Available widely enough that it can be subjected to Machine Learning procedures.
FMI: Finnish Meteorological Institute
Gabor filter: a texture feature, detects edges
GIS: geographic information systems
GPR: Ground Penetrating Radar
grid constant: grid or raster size, pixel size. The pixel has usually a square shape.
GTK: Geological Survey of Finland
hydraulic conductivity: speed of water flowing downwards through soil layers. Related to grain size distribution of soils. (Alternative form in one reference: water permeability/conductivity)
Inspire: EU directive 2007: Infrastructure for Spatial Information in the European Community, see http://inspire.ec.europa.eu/
LBP: Local Binary Patterns, a texture feature, detects patterns
LiDAR: Light Detection And Ranging, an ALS technology. In this project only airborne LiDAR (ALS) data is used
LUKE: Natural Resources Institute Finland, from 2015 (Luonnonvarakeskus in Finnish)
Metla: Finnish Forest Research Institute
MLP: Multilayer perceptron, one neural networks formulation
MS-NFI: Multi-Source National Forest Inventory (VMI in Finnish) by Metla
NLS: National Land Survey of Finland
Peatland mask: a binary-valued nation-wide raster map indicating the peatland areas
PSI: EU directive 2003: Public Sector Information re-use
Primary feature: original remote sensing or field observation
Remote data: airborne or satellite data, can be passively (optical remote sensing) or actively (Radar, LiDAR) sensed
RSAD: Radar Surface Arrival Detection, a technique to determine ground surface dielectric permittivity by measuring the travelling time of a radar signal with a ground penetrating radar by keeping the distance between a transmitting and receiving antennae constant
SAR: Synthetic Aperture Radar, a special arrangement to implement a compact satellite receiver antenna and the corresponding computational tas and the data format
TDR: Tilt derivative is the arctangent of the ratio of a vertical to a combined horizontal derivative.
Trafficability: The ability of terrain or road to support the passage of vehicles.
UTM: Universal Transverse Mercator

1 OVERVIEW OF THE PROJECT

There is a global trend towards increased availability of governmentally collected information. The EU Inspire directive has set specific goals and a timetable for this process. Another similar directive is EU directive 2003: Public Sector Information re-use (PSI). While commercial applications based on location data already exist for urban areas, open data applications utilising environmental data are only just emerging. There are excellent data collections in Finland, which if used judiciously, could boost economic activity and produce new business models, or, on the other hand, could be used to prevent economic damage caused by e.g. flooding or storms.

Thus far these resources have been available to only a small group of experts and it was not possible to gauge the added value of a unified treatment of these data sources, either in theory or practice. Now when this data is publicly available, one can expect new results related to the location and utilisation of natural resources, as well as an improved understanding of both the many kinds of phenomena involved and of geographically related multidisciplinary topics. Moreover, new European satellite programs such as Sentinel are becoming operational, delivering virtually real-time remote sensing, free of charge.

An executive summary of the project is provided at Sec. 1.4. The mission of this project was to:

1. Demonstrate the usefulness of currently or imminently available open data with pilot applications
2. Test current data management platforms and data accessibility via pilot applications
3. Method development for data understanding (visualisation)
4. Method development for combining multisource (heterogenous) data for analysis purposes (e.g. clustering, regression, classification and time-series analysis)

The actions taken to reach these goals were:

1. Analysis of databases and data formats: data currently exists in different formats and a reasonable policy had to be outlined to facilitate efficient analysis
2. Selection of data handling platforms: Survey and selection of one or more of the existing platforms
3. Selection of test areas for the pilot applications
4. Data collection and calibration
5. Data analysis using primarily well known Machine Learning methods: clustering, regression, classification, time-series analysis
6. Pilot applications and visualisation: three pilot applications for open data usefulness demonstration and testing of developed methods, with specific focus on those methods which facilitate spatially partitioned handling of massive data.
7. Finding future applications and new commercial possibilities (business potential). Emerging cloud technology is evaluated in this respect.

1.1 Project participants and their data resources

The four participants each have strong natural resource and environmental data capabilities and there is potential for synergy gain with the use of well-selected applications. The following is a list of the participants and their nationwide resources. Each party has case and area specific data collections:

- Finnish Forest Research Institute (Metla): The thematic forest map of the Multi-Source National Forest Inventory (MS-NFI), 43 attributes describing the forest state, the product is updated every two years (2009, 2011..).
- Geological Survey of Finland (GTK): Airborne low-altitude geophysical data (gamma-ray surveys, electromagnetics), Quaternary sediment mapping data

- Finnish Meteorological Institute (FMI): An abundance of weather and ground related data from both meteorological and satellite sources.
- University of Turku (UTU): machine learning and image processing experience.

Geographical/topographical data from the National Land Survey of Finland (NLS/MML) was also used, although NLS was not a participant of this project.

In addition to these existing nationwide data collections, field observations and when necessary, limited new field campaigns were applied for obtaining the ground truth data needed for development and validation of the applications.

1.2 Goals

The main aim of this project was to develop and adapt existing computational methods of specialised massive data analysis and to find practical information for use in a variety of applications. Another aim was to develop methodologies for data mining which allow effective analysis of current data sources and easy access via open interfaces so that the results can be utilised by as many as possible. Emerging cloud technologies were evaluated for this purpose.

The planned final outcome of this project was to enable new commercial activities based on public data, keeping in mind the needs of both traditional industry as well as small and medium size enterprises.

The following is a list of the goals of our partners.

Terrain trafficability in timber harvesting (responsible partner Metla)

Nowadays the uninterrupted raw material chain from the forests to the mills is of central and upmost importance to the forest industry. Difficulties are caused above all by poor soil load bearing capacity mostly due to melting soil frost in the spring and ample rainfall in the autumn. The goal was to develop an approach for estimating the terrain trafficability in timber harvesting based e.g. on weather conditions (history, forecasts), snow, and frost data combined with soil and forest information. The applicability of real time telemetric information from harvesting machinery should also be considered.

The developed approach should enable near real-time trafficability maps. Trafficability maps have been mentioned by the forest industry as one of the potentially important tools for timber harvesting and secondary transportation.

Arctic Infrastructure (responsible partner GTK)

Due to future challenges associated with global change and increasing transportation demands, particularly for the mining industry in the arctic/subarctic, effective tools are needed to outline subgrade/soil quality and aggregates for road/railroad networks as well as for subgrade bearing capacity.

The goal was to reveal construction aggregates with the integration of airborne gamma ray and Airborne Laser Scanning (ALS) data. The focus was on identifying coarse-textured hummocky moraines (so-called mass-flow deposits) for construction aggregates instead of commonly used glaciofluvial materials.

Detection of environmental hazards by applying radar data (responsible partner FMI)

The aim of the project was to investigate and demonstrate the utilisation of satellite (remote sensing) observations for applications related to timber harvesting (monitoring of soil frost status) and environmental disasters, such as the monitoring of flooded areas and storm damage to forests.

Concerning both applications, assessment of the feasibility of novel satellite observation systems is the main topic. FMI operates the Sodankylä satellite data centre which has started the acquisition and real-time processing/delivery of satellite data products relevant to these issues. In particular, the Italian COSMO-SkyMed synthetic aperture radar (SAR) satellite constellation system has high potential for near real-time disaster monitoring. Sentinel-I SAR satellite data was not yet available for this project, but it is considered a future possibility in this area of research. It will provide a large number of image products thereby enabling advanced monitoring of environmental changes, e.g. the detection of fallen trees using high density image products.

1.3 Project timetable and chronological outline of project activities

Table. 1.1. Project timetable and main tasks.

Action	Organization	2013				2014				person months TEKES	person months Own funding
1) Analysis of databases and data formats	TY	■	■							1	0,25
	GTK	■									
	IL	■									
	Metla	■									
2) Selection of data handling platform	TY		■	■	■					1	0,5
	GTK		■								
	IL		■								
	Metla		■								
3) Selecion of test areas	TY	■								0,5	0,25
	GTK	■									
	IL	■									
	Metla	■									
4) Data collection and calibration	TY		■	■	■	■				4	1
	GTK		■	■	■						
	IL		■	■	■						
	Metla		■	■	■						
5) Algorithms and analysis	TY			■	■	■	■			10,5	5,5
	GTK					■					
	IL					■					
	Metla					■					
6) Pilot applications and visualization	TY			■	■	■	■	■	■	4	3
	GTK			■	■	■	■				
	IL			■	■	■	■				
	Metla			■	■	■	■				
7) Future applications and new commercial possibilities	TY								■	1	0,5
	GTK								■		
	IL								■		
	Metla								■		
Total									■	22	11

The following provides an expanded list of the project phases 1) - 7). It is also a cursory presentation of the results:

1.3.1 Analysis of databases and data formats

An initial plan of hierarchical data storage was formed. Several alternative data formats were tested, including multilayer geoTiff images, the so called .xyz format and various raster data files. Since computation is mainly of the Machine Learning (ML) genre, the usual GIS approach was abandoned. The final format is a generalisation of the .xyz format discussed in Ch. 3. The final format is suitable for independent analysis software and algorithms.

A simple elimination scheme was used to filter out the non-meaningful entries. This was due to the impressive quality of the data (each case contained at least 90% valid entries).

The raster materials present many challenges due to the following reasons:

- Massive amount of information
- Unstructured nature of some data
- Temporal and spatial heterogeneity
- Lack of unification resulting in alternative formats
- Difficulties transmitting and re-transforming the data
- Numerical features are not directly usable

- The reliable interpretation of biophysical characteristics requires a combination of multiple data sources and finely tuned numerical analysis

1.3.2 Data handling platform

Most of the analysis scripts are simple and can be implemented in either Python or c language, the latter via automatic translation. The real problem is a descriptive top layer for a collection of scripts. Several alternatives were inspected. A visualisation demonstration was based on Amazon EC2.

The technology is rapidly transitioning and new opportunities are emerging. Future applications could be bundled with services like www.paikkatietoikkuna.fi or cloud integration methodologies currently developed at Probis and Techila, GIS end-user applications developed at GIS Cloud and so on. The on-going march of these new possibilities directed us to cancel the platform design.

The concrete deliverables of this project have been proposed to be included in the FMI lead national satellite data center under development in Sodankylä, see Sec. 5.4. Other potential host technologies are addressed in more depth at Ch. 5.

1.3.3 Selection of test areas

Forest and soil types are highly variable in Southern Finland. Both Parkano and Pieksämäki facilitated useful measurements. The areas have more internal variety, useful test fields and available data than was required in this project.

Northern Finland also has plenty of high quality measurement series and test sites. Pomokaira in Sodankylä and Kemijärvi were selected as test sites in Northern Finland. Pomokaira represent a target area were subgrade soils are composed mostly of fine-grained tills therefore being a suitable for studies of bearing capacity. Kemijärvi represents the largest hummocky moraine field in Finland. Therefore it is a perfect target area for mapping of potential coarse aggregate materials for construction purposes.

Moreover, additional fieldwork was conducted by Metla in Evo, Pieksämäki and Kolari study areas during 2014.

1.3.4 Data collection

Data collection required some field efforts from Metla and GTK. The field campaigns included:

1. timber harvesting machinery rut formation and soil penetration resistance measurements at Pieksämäki,
2. flood reference in Evo and Kolari and
3. drainage network references in Evo.
4. particle size data by GTK in Pomokaira
5. ground penetrating radar (GPR) measurements by GTK in Pomokaira

Also, several minor changes had to be made to the data acquisition plans during the project. FMI used the COSMO-SkyMed satellite with Synthetic Aperture radar (SAR) for flood detection.

There were four campaign areas with whole area covered by public data. These were Parkano, Pomokaira, Kemijärvi and Pieksämäki. Evo and Kolari were covered by geographic data of NSL and by remote data (aerial and satellite). There were several other areas, which were excluded from this project, but which would be useful when the water budget modelling is addressed by the multi-source data and Machine Learning approach used in this project.

1.3.5 Algorithms

The algorithms included ridge regression, k-Nearest Neighbors (k-NN), multi-layer perceptron (MLP), Self-Organising Maps (SOM), feature selection and k-cross validation with dead zone, see Sec. 3.3. Also several imputation methods and derived feature generation algorithms were tested. These included some ad-hoc methods, texture methods and traditional geographical methods.

1.3.6 Pilot applications

Pilot applications included test cases using Machine Learning analysis and a demonstration of the Amazon EC3 platform and GUI using hierarchical raster files.

The test cases are listed below (Table 1.2). Only one aspect of a possible application has been addressed. Application sectors are: soil model and trafficability, soil model and arctic infrastructure, flood detection and forest drainage network evaluation:

Table 1.2. Project cases by task type, site and calendar date.

subject	task type	area of interest	application	time period
Soil type	prediction	Parkano	soil model	10-12/2013
Soil hydraulic conductivity	prediction	Pomokaira	soil model	04-06/2014
Flood state	detection	Evo, Kolari, Kittilä	flood detection	05-07/2014
Drainage network	detection, evaluation	Evo	extent mapping, condition assessment	08-12/2014
Soil damage	prediction	Pieksämäki	soil model	02-09/2014
Sand deposits	prediction	Parkano	(arctic) infrastructure	10/2013
Ground penetration	prediction	Pieksämäki	soil model	10-11/2014
Feasibility for generalisation	estimation	Parkano	general	12/2014
Load bearing capacity of forest roads	prediction	Pomokaira	arctic infrastructure	12/2014
Mass-flow deposits	prediction	Kemijärvi	arctic infrastructure	08,12/2014

1.4 Executive Summary

The highlights of this project were_the successful mapping of forest floods, a comprehensive mapping of drainage networks based on airborne laser scanning, hydrological model HOPS, geomorphological recognition of mass-flow deposits and good prediction performance of forest terrain trafficability and subgrade bearing capacity. Further details regarding each case study are provided in sect. 4.

The terrain trafficability prediction results were promising, however further development is needed, i.e. including sensing equipment into harvesting machinery for continuous online field data. It was obvious that the requirements for field data are high and streaming measurements will be required for a practical application. The first Pieksämäki data set indicated that coarse strip road soil damage classification cannot be used for damage modelling without knowing the extent of vehicular loading at each location.

The application of more accurately measured field data increased estimation potential notably, yet collecting it manually is expensive. The next logical step towards an operational application is to instrument forest harvesting machinery to collect terrain trafficability and damage data during harvesting operations. This could be achieved by measuring the power consumption of the machine transmission via CAN-bus and linking it with physical soil damages by modelling based on laser scanning and auxiliary data sources. The field measurement stream and generic spot wise predictions are then combined in situ to estimate the trafficability of the surrounding terrain. The estimation accuracy improves over time when new information is cumulating.

The use of COSMO-SkyMed SAR data to detect flooded areas in forests and open areas was successful. Compared to traditional methods, SAR based flood analysis is very fast and cost efficient. Therefore the retrieved flood maps can be beneficial for a variety of users such as emergency personnel, insurance companies, landowners and farmers, as well as private and public institutions for community planning. This application could also be potential for more accurate estimation of the progress of the flood development during the spring snow melting period.

Radar imagery is not sensitive to weather conditions making it a promising tool for storm damage observation campaigns on large forest areas. The detection of storm damages was studied on areas where individual trees were felled to create simulated damage. The results using X-band radar imagery did not look promising. Obviously there was too much noise which prevented the detection of the damages from a high energy SAR radar imagery. An L or C band radar, such as Sentinel-I could be more potential for detection of storm damages.

Ground penetrating radar (GPR) was used to classifying the bearing capacity of the forest road network particularly during the period of spring thaw weakening. It is an economic alternative to soil textural analyses, and may prove a tool for assisting nationwide road transportation forecast system.

Hydrologic prediction system (FMI HOPS) is a ready off-the-shelf result of the project. HOPS has the usage of its own in many fields, and it is the first finished component of a possible future trafficability forecasting system. It has a built-in property of keeping track of precipitation and evapotranspiration. While the calculation grid is sparse, it can be used to provide the dynamical weather dependent signal to water budget model. The water budget model is not a project deliverable, but preliminary studies have been made to properly understand the Machine Learning aspect of the future water budget models.

The application on the use of airborne laser scanning for the detection of the forest drainage network worked well in the Evo test area. Even from the low pulse density laser data it was possible to locate the ditches more accurately than currently on base maps and at the NLS topographic database. Proper and accurate drainage models will be essential when including the moisture models into the forest applications presented in this work and when practical applications will be developed. One important application can be found in assisting drainage network maintenance planning by estimating the condition of the network to the level of an individual ditch. This could benefit forest companies and owners. As there are millions of hectares of drained land in Finland alone, the results potentially have national scale importance.

Application of the ALS data for detecting the mass-flow deposits suitable for road and other infrastructure construction is promising. Especially, in the Arctic regions this application is particularly important in locating proper construction materials in order to minimise the environmental impacts.

There seems to be not many similar Machine Learning projects using the multi-source locational data to such multiple goals. We did not find general methods to estimate the generalizability i.e. the spatial extend of the prediction performance. This pointed us to develop a new dead zone validation methodology.

The best prediction performance was met with a path selection test, which gave good performance indicating that the currently available public data is enough for predicting forestry trafficability what comes to route selection aspect.

The project participants share the view that a public data storage solution is needed. It would help the research projects and communication across organisations and could be a base for many commercial projects.

1.4.1 Utility value of various data sources

The following Table 1.3 summarises the relative importance of all the data sources in the test cases of the project. The data source can be useful even when only one of the derived features based on this source is useful. The data sources are considered as follows: ** High importance, * medium importance, - no importance, + potential importance (not tested), ++ high potential importance (not tested) and *+ mediocre importance , additional potential exists. x means input feature, which is the subject of prediction.

Table 1.3. Estimated importance of data in each case study. The data sources are considered as follows: ** High importance, * medium importance, - no importance, + potential importance (not tested), ++ high potential importance (not tested) and *+ intermediate importance, additionalpotential exists.

DATA	Availability	4.1.1, Soil type	4.1.2, Hydraulic conductivity	4.1.3, Forest roads	4.1.4, Soil damage 1	4.1.5, Soil damage 2	4.1.6, Route selection	4.2, Forest flood	4.3 Drainage network	4.4, HOPS	4.5 Storm detection	4.6 Mass-Flow
LR ALS (NLS)	Public		*	+	*+	*+	*+	**	**			*
HR ALS	Closed		+	+	++	++	++	++	**			++
MS-NFI	Public	**	*		*	**	**	*				+
FMI-Weather	Public			++	**	++	++					
GTK-Soil texture	Public		*	+	*	**	*					+
GTK-GAMMA	Public	*	**	+	**	*	*					+
GTK AE	Public	+										
COSMO-Skymed (SAR)	Closed	*	+	+	-	-	-	**			-	
Sentinel-I (SAR)	Public (11/2014)	++	+	+	+	+	+	++			+	
Field data applied	-				yes	yes	yes	yes	yes		yes	yes

Data of three test case areas (Parkano, Pomokaira, Pieksämäki) was preliminarily analysed by feature selection. There were no clearly dominant data sources and each one contributed to the final performance. At the moment we can omit appr. 30-40% of the 60-90 features used in each case, but unnecessary features fall into all data source categories. It seems that the selected cases have different demands and each data source is useful in some cases.

Section 6.4.1 lists possible future applications based on a generic data platform providing an access to these data sources. Bi-weekly updated trafficability forecasts with local details on national level for both forestry machines and forest roads is a possibility, yet considerable research and some industry agreements have to be made first. It is important to notice that available remote sensing data allows us to generalise field detected trafficability. However, stand alone estimates from remote sensing may be beyond the presently available remote sensing data. Combining the wide-area automated geomorphological classification with additional features derived from airborne laser scanning (ALS), one can pinpoint e.g. mass flow deposits with infrastructure value. This application has potential for an international product. Flood detection is highly useful product, which could be finalized as a stand-alone version, but which would be logical as a part of a generic data platform. The remote assessment of the condition of forest drainage network have many applications, it can e.g. improve the performance of the trafficability applications but also help with drainage network maintenance planning, this highlights the importance of ALS data. Higher pulse density of ALS data would enhance the target signal, e.g. ditch or mass-flow deposits, and is therefore hoped to be available in the future as an open data set. It could also open completely new economical possibilities.

The common opinion of the project participants is that there are still some useful applications to be found, and that the performance of current test cases can be improved by better methods. An integrated data platform would speed up the cycle of research and innovation considerably. It is noticeable that when open data is accumulating and

it's potential is thoroughly researched in the future, it is considered to create huge potential to discover and develop even more and different applications. This could be one of the key factors when building efficient and economic systems for both industry and public services. This project included only few potential application areas and further research and opening new data would almost certainly be beneficial in the future.

2 DATA

This Chapter tries to address each field of expertise of participants, and the intended prediction tasks. Ch. 6 contains a similar treatise about the possible existing and future data sources not used in this project.

2.1 Public open data

This is a short introduction to the data used in the project. Data consists of metadata, which identifies the time and coordinate frame, as well as its features. A feature of this presentation is a numerical aspect applicable to Machine Learning methods and available throughout the data samples. A detailed list of all features can be found in the Appendix.

2.1.1 Features by data provider

The following is a rough categorisation of approximately 50-60 primary features used in the project. The features are grouped by data provider. FMI refers to the Finnish Meterological Institute, GTK to the Geological Survey of Finland, NLS to the National Land Survey of Finland and Metla to the Finnish Forest Research Institute:

FMI: Frequently updated data which can be utilised when forecasting terrain trafficability. Raster size varies between 2 m … 200 km. See the open data site at: https://en.ilmatieteenlaitos.fi/open-data

- topsoil humidity available from Sentinel-1 Synthetic Aperture Radar (SAR) on C-band (~3 cm wavelength) since the end of 2014.
- An abundance of weather related data from both meteorological and satellite sources, as well as simulated soil moisture and soil temperature data, including merged data products based on satellite data assimilation.
- estimated precipitation and evaporation, effective temperature
- Cosmo Skymed SAR (till the end of 2014, non-open product) and Sentinel-1 SAR (from Nov 2014, open product)

GTK:

- Basal and surficial sediment mapping data, potassium window data from airborne gamma-ray surveys, apparent resistivity data from airborne electromagnetic surveys, see Sec. 2.3
- RSAD field measurements with ground penetrating radar (Pomokaira) and analyses of particle grain size distribution of sediment samples (Pomokaira)

NLS: See the site at: https://tiedostopalvelu.maanmittauslaitos.fi/tp/kartta?lang=en

- topography, road network (.asc, .png and ESRI shape)
- topographic database (ESRI shape)
- e.g. peatland area, water area, drainage network, roads
- Airborne Laser Scanning (ALS) data (.las and .xyz -format), see Sec. 2.4

Metla: Case study field measurements and (MS-NFI), which is updated once every two years (2009, 2011 ...) and is in GeoTIFF format.

- Multi-Source National Forest Inventory, MS-NFI, see Sec. 2.2

Field-measured features are:

- estimated harvesting site damage class (data was kindly provided by **Metsäteho Oy**)
- harvesting machinery rut formation (Pieksämäki)
- soil strength as described by the penetration resistance (Pieksämäki)
- field recorded drainage networks (Evo)
- field reference for forest flood extent in Evo and Kolari research areas

The scope and qualities of these open data are essential to this project, thus the data aspect will be discussed in other chapters too:

- Ch. 4, methods used to produce the data, its connection to expert fields in question and to phenomena to be predicted.
- Ch. 3 and 5, data preprocessing and cloud arrangements, possible additional data
- Appendix A1: Technical listing of data features

The raw data, after the initial spatial cut, some pre-processing (elimination and/or imputation of missing values, interpolating to new grid sizes, regularisation, normalisation etc.) yields the primary features. The set of primary features was dictated by practise in each case.

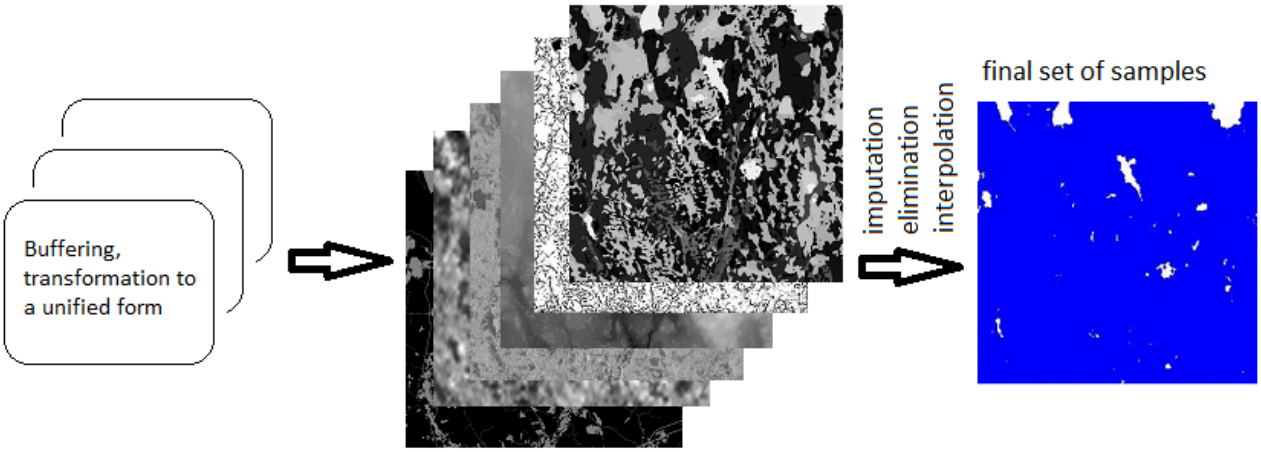

Fig. 2.1. Workflow: Spatial selection, transformation and assembly and pruning the data. Figure by P. Nevalainen, UTU.

2.2 The forest inventory (MS-NFI, Metla)

National Forest Inventory (NFI) is sampling-based inventory system maintained by the Finnish Forest Research Institute (Metla). The NFI is a sampling-based inventory system that covers all land-use classes and ownership categories throughout the whole country. Main purpose of NFI is to produce reliable information on forest resources and growth, the health of forests, forest biodiversity, and future cutting possibilities at national and regional forest level.

Multi-source national forest inventory (MS-NFI) is a method developed for producing estimates of forest resources for areas smaller than regional level and, in the form of thematic digital maps. The satellite image based MS-NFI was introduced during the 8th NFI at the end of 1980's. The method employs satellite images, digital maps and NFI field measurements as data sources. Mainly high resolution satellite images, such as Landsat-series (TM,ETM+, OLI), IRS and SPOT, have been applied in MS-NFI. NLS topographic database is used to separate forested areas from other land-cover categories. Furthermore, stratification of forests into peatland and mineral land strata is carried out based on NLS topographic database. A digital terrain model is used to reduce image spectral distortion effects caused by topography, and also for vertical stratification.

Non-parametric k-Nearest Neighbor (k-NN) estimation method is used in MS-NFI calculations for deriving forest estimates of individual satellite image pixels on the basis of NFI field plots as reference data. The MS-NFI thematic maps are raster format digital maps with a spatial resolution of 20 m. MS-NFI data contain over 40 forest variables in the form of thematic maps, including, e.g., the volumes by tree species and timber assortments, stand mean variables, the biomass by tree species groups and tree compartments and forest site type characteristics. Estimates produced in MS-NFI cover in principle 100% of the land area, but there might be some missing regions, for example because of cloud coverage.

2.3 Airborne geophysical data and Quaternary geological mapping data

Finland has been covered with systemically collected airborne data by the the Geological Survey of Finland. The airborne geophysical surveys was carried out since 1972 till 2007 with a 30 m nominal flight altitude and 200 meter line spacing. The geophysical parameters measured are the Earth's magnetic field, the electromagnetic field and natural gamma radiation.

Magnetic measurements determine the Earth's magnetic field strength (magnetic flux), obtaining the total magnetic intensity of the Earth's magnetic field as a parameter. Magnetic data is applied to ore exploration, bedrock mapping and environmental studies at the regional scale.

Airborne gamma-ray measurements register the Earth's natural gamma-ray radiation reaching to the depth of less than 1 meter and thus provide an effective alternative for soils mapping.

Airborne electromagnetic measurements obtain conductivity information from the ground with a maximum penetration depth of 70-100 m. The apparent resistivity, which was calculated from the real and imaginary components of the electromagnetic field, is a good tool for locating conductivity structures in near-surface geological and environmental applications.

Systematic mapping of Quaternary deposits in the scale of 1:20 000 has been carried out by the Geological Survey of Finland since 1870 till 2010. The data represent distribution of organic and inorganic sediments and their geomorphological origin as polygons over the landscape in a GIS vector dataset. The mapping consists of visual pre-interpretation of aerial photographs and field observations for the map interpretation. The sediments are classified into organic deposits (mud, sedge and Sphagnum peat), unsorted glacial mineral sediments (gravelly, sandy and fine-grained till), sorted coarse grained sediments (boulders, gravel, sands), and sorted fine-grained mineral sediments (silts, clay). In addition outcrops and shallow bedrock regions are mapped. The primary polygons represent the basal sediment at the depth of 1 meters (minimum polygons size 2 ha) which are overlain by superficial sediments (0.4-0.9 m, minimum polygons size 4 ha). Data is public, open and free of charge available at http://hakku.gtk.fi/. The data covers approximately one third of the country, especially the terrain in the southern part of Finland and around the largest cities is mapped.

2.4 Airborne Laser Scanning data and NLS topographic database

The NLS Topographic database contains the basic elements which are used in map making. Its key objects on the map are, for example, the traffic route networks, buildings and constructions, the administrative borders, geographic names, land use, and elevation contours. It also depicts features such as peatland areas, forest areas and drainage networks, which were utilised in this project. The positional accuracy of the Topographic database corresponds to that of scales 1:5 000 - 1:10 000. The traffic road network and geographic names are updated constantly, buildings, constructions and administrative borders annually and the other elements approximately every 5–10 years (NLS 2014). A thoroughly description about the product is available at http://www.maanmittauslaitos.fi/en/digituotteet/topographic-database

Two types of Airborne Laser Scanning (ALS) data were used. Primary data source was the publicly open ALS data of the NLS. This data is expected to cover the entire Finland by the year 2019. The data is distributed in two formats, 1) point cloud (.laz) and 2) rasterized Digital Elevation Model (DEM). The point cloud has appr. 0.5 pulses per square meter, and it is therefore considered low pulse density data. The resolution of the NLS DEM produced from the ALS data is 2 meter and its vertical accuracy is estimated to be 0.3 m. NLS ALS product will cover the whole Finland by the end of 2019, but possibly earlier.

In addition, we tested high pulse density data in Evo for the drainage network case study in order to assess the effect of pulse density to the results. The high density ALS data were acquired by FM-International Oy in July 2009 using a Leica ALS50-II SN058 laser scanner. The scanning was carried out from a helicopter flying at the altitude of 400 m above ground. The lidar footprint of high density data was 10 cm and pulse density 10.43 pulses/m2. The ALS data were geo-referenced in the EUREF-FIN(ETRS89) coordinate system.

All of this NLS data is distributed free-of-charge under open data licence described here: http://www.maanmittauslaitos.fi/en/NLS_open_data_licence_version1_20120501

2.5 COSMO-SkyMed SAR data

COSMO-SkyMed satellite data was used in flood mapping and in the forest damage demonstration. The COSMO-SkyMed constellation consists of four individual satellites equipped with X-band Synthetic Aperture radar (SAR) sensors. COSMO-SkyMed is ideal for mapping natural hazards and damage, because it enables daily imaging of any region on the earth. COSMO-SkyMed offers several imaging modes, where the spatial resolution ranges between 0.5 m and 100 km. For higher spatial resolution, the image size is smaller. The new Sentinel-1 SAR satellite has become functional on Nov 2014, and will also provide data in the future.

3 PREDICTION PERFORMANCE

There are two seemingly contradictory approaches to predicting geographically distributed entities. One is traditional spatial interpolation, the other is Machine Learning prediction with generalisation in new areas.

For the first approach, we use the term 'interpolation' as a general label since the division between interpolation and extrapolation is not always clear as it depends on the scale used and accuracy required. Spatial interpolation uses known property values at certain reference points and estimates the expected value field for new points. Since reference data is usually randomly scattered, practical implementations use various radial functions. Spatial correlation indicators, like variogram are used to determine the best radial scaling, either locally or globally. A modern approach is surrogate modeling, where incomplete field variable distribution knowledge is used to form the expected error term, which is then minimised using performance validation methods similar to Machine Learning in order to fix the free parameters.

The Machine Learning approach used in this project is very similar to interpolation. A portion of the given data is used as a verification set. This project has also unearthed additional challenges, for example, generalisability in new areas which have no field measurement data. Generalisability can be defined as having both complete spatial independence and high prediction performance at the same time. Texture categorisation and segmentation is another challenge to be addressed.

The problem in achieving the goal of reliable generalisation basically has three key aspects:

- Is the given data set representative enough? Experts can only provide a partial answer, absolute certainty would require an expensive and perhaps impossible data gathering campaign.
- How far from the initial domain prediction can it be extended? This can be partially evaluated by methods developed in this project, e.g. performing dead zone analysis by producing the prediction performance curve as a function of the distance from the nearest known measurement. See e.g. Fig. 4.3, where the 'features only' curve in the right plot settles to concordance index (CI) of 0.55 at appr. r= 200m. CI=0.5 means completely random choice.
- What is the importance of pairing the used feature set with location data? In one extreme location has no importance and the dead zone performance curve is steady, in the other extreme all features are meaningless, and the classical spatial interpolation is the only tool.

To better understand the generalisation problem, a short formulation of the prediction problem follows in sections 3.1-3.3.

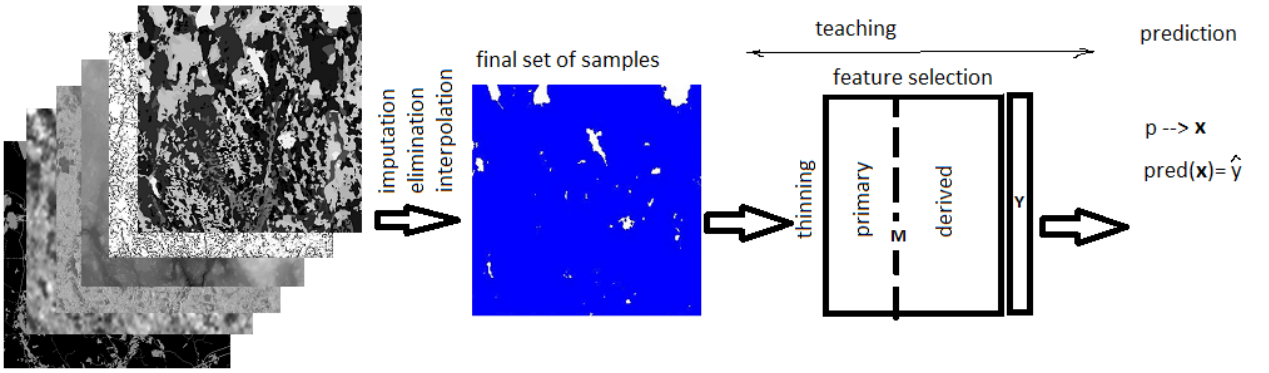

Fig. 3.1. Workflow: Initial data matrix assembly, forming the data matrix M, reduction of the size of the problem, teaching and delivered predictor. Figure by P. Nevalainen, UTU.

3.1 Data matrix and prediction

In GIS the location is a central concept, and this is reflected on the code level, too. Most computational algorithms integrated into GIS software are laid down in a pixel or grid oriented fashion similar to image processing algorithms.

Data can be either organised or randomly located, and the final formulation has each location as one row of the data matrix. The metric relation between samples can be based on the features only, location only or to a mixture of features and location. Teaching based on location only resembles the interpolation paradigm. In this case it is a variant of interpolation. The following is a short presentation of the initial data handling process.

The raw data, after some pre-processing (elimination and/or imputation of missing values, interpolating to new grid sizes, scale transformation, regularisation, normalisation etc.) yields the primary features. The set of primary features is dictated by practise in each case. The field is open in order to find more features such as satellite images in the visible and infrared ranges.

3.1.1 Derived features

There are approx. 40 derived features, which are based on the primary ones (see Ch. 1.5.1 and Fig. 2.1). The purpose of the derived features is to improve prediction performance by analysing local texture and local variance etc. In most Machine Learning methods the derived features usually add little to the computational load. Thus, adding one new derived feature can be judged on a case by case basis using the performance test. Derived features are computed using the following methods:

- local averaging, local gradient, Gabor filtering (applied to all continuous feature fields)
- texture methods: local binary patterns (LBP), local mean and variance (3x3...11x11 windows)
- large texture windows, large scale sequencing of geomorphology (these two methods are not employed in this project).

The benefit of each derived feature was evaluated afterwards by performing a prediction test, both with and without the feature in question. The subgroups of derived features were not tested. As an example of derived features is topographic height, which, as a primary feature, brings forth a wide sortiment of derived features, see e.g. Topo Toolbox (link) about the typical topography based derived features.

A meaningful derived feature usually represents a physical property which has an effect on prediction performance. A typical example is localised topological height. It is likely that having an indicator for small hills is useful, and in this respect the actual choice of the filter formulation used to produce the local height values is not so important. We may have omitted several useful derived features, but this field remains open for experimentation. There are also blind trials for finding derived features, including local mean, local variance and various texture features. If they do not come with a high computational cost and improve the prediction performance when applied. For further information on feature selection, see Kohavi (1997) and Blum & Langley (1997).

3.1.2 Data matrix structure

The public data is typically regular grid data interrupted with occasional missing values. Field measurements are totally irregular observations. Both deserve a unified representation for computations presented in Eq. 3.1. Data matrix structure is very standard to the Machine Learning approach. An individual observation or public data grid point forms a row in data matrix **M**. The columns are:

- raster indices *i* and *j*. These are usually local, and they are recorded to speed up some neighborhood calculations
- location (x,y) in ETRS-TM35FIN coordinates in meters in East, North order
- features *fij,i = 1…n,j = 1..m* where *n* is number of samples and *m* is number of features
- target features *ti, i = 1..n*

$$\mathbf{M} = \begin{pmatrix} i_1 & j_1 & x_1 & y_1 & f_{11} & \dots & f_{1m} & t_1 \\ i_2 & j_2 & x_2 & y_2 & f_{21} & \dots & f_{2m} & t_2 \\ & & \vdots & & & \ddots & \vdots & \\ i_n & j_n & x_n & y_n & f_{n1} & \dots & f_{nm} & t_n \end{pmatrix} = \begin{pmatrix} \mathbf{M}_1 \\ \mathbf{M}_2 \\ \mathbf{M}_3 \end{pmatrix} \quad (3.1)$$

The matrix can be dissected in vertical direction to three parts, which represent different sample subsets for training, validation and testing:

$$\mathbf{M}_i = (\mathbf{X}_i \mathbf{Y}_i),\ i = 1,2,3\ (3.2) \quad (3.2)$$

X is the feature matrix and **Y** is column vector of the target features. There is a version of matrix file **M** in Eq. (3.1) without the grid indices (first two columns). It is used with the scattered measurement points. That version is the most common in this project, only one case area (Parkano) had almost full raster coverage for both **X** and **Y** features and the index columns **i** and **j**. Almost full raster has many advantages, e.g. the neighbourhood sets *Nr* with radius *r* can be computed in *0(|Nr|)* time.

The matrix structures (Eqs. 3.1 and 3.2) are presented, since they drastically differ from the usual approach of algorithms integrated to existing GIS software. Since problems can be huge, matrix **M** has to be generated column by column from GIS platforms and the result saved outside the GIS before the Machine Learning codes can start the analysis. This is reasonable, since a large set of ready-made software accepts matrix **M** format, while only few integrate readily to GIS and those that do generally cause memory issues.

3.1.3 Training and prediction

The Machine Learning methods used (ridge regression, k-NN, MLP) can be expressed as a triplet *(Train,Pred,V)*, where *Train* is the training, *Pred* the prediction operator and *V* is the loss function. Operators *Train* and *Pred* are paired by the state *S* which *Train* produces and *Pred* uses. Typically data **M** is divided into three subsets, two of which are addressed in Eq. (3.2). The best possible state *S* will be chosen by means of optimising total error *e12*:

$$S = Train(\mathbf{X}_1, \mathbf{Y}_1), \widehat{\mathbf{Y}}_2$$
$$= Pred(S, \mathbf{X}_2),\ e_{12} = V(\widehat{\mathbf{Y}}_2 - \mathbf{Y}_2) \quad (3.3)$$

where *V(.)* is the loss function (cost of error). This arrangement is enough to find the best model for part 2 of the data set. In practice there are several methods *(Train,Pred)* to be compared and those methods should also be compared using new data not previously included in the train-validate phase. Therefore a train-validate-test scheme is adopted. The best state *S* found is subjected to a prediction in the third portion *i=3* of data not used yet. This arrangement simulates the generalisation situation and is a good basis for comparison between different methods. Since the test data (*i=3*) is meant for comparisons between methods, it plays no part in actual methodological part and is not discussed further.

Since the presentation would be rather involved with validation schemes and possible dead zone alteration of the train-validate phase, a train-validate-test operator *A* is introduced:

$$e_3 = A(Train, Pred, \mathbf{X}, \mathbf{Y}) \quad (3.4)$$

k-fold cross validation is a common scheme, which reuses the data in selecting train-validate sets. The operator A may alter the basic k-fold cross validation scheme in many ways. Leave-one-out (LOO) scheme results by choice $k=n$, where each n validation sets have only one point. The dead zone approach removes points creating an extra margin (dead zone radius) between the train data and validation data. This is done in order to verify the effect of spatial correlation in prediction. There are many alternatives for the dead zone approach. Data can be artificially divided into two distinctive and separate sets, or the LOO scheme can be used. Fig. 3.2. represents these variations in a simplified visual form.

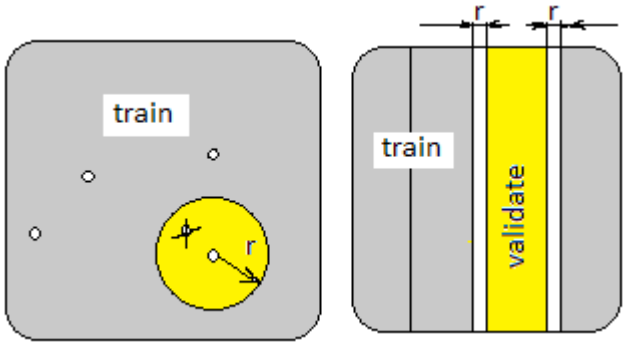

Fig. 3.2. Dead zone condition for train and validate sets: no training sample and validation sample is closer than dead zone radius r. Sets can still be spatially mixed (not depicted), separate (k-fold cross validation, right), or LOO (left). Figure by P. Nevalainen, UTU.

3.1.4 Dead zone validation

The dead zone method modifies the training set in relation to the validation set so that teaching utilises no information from closer than a given radius r. This simulates the problem where a new prediction point is at a distance of r from the nearest known data points. Possible arrangements for this are many, the following is a list of some possibilities:

- Leave-one-out (LOO) dead zone validation has validation points pruned out by the training set until the dead zone condition holds, see Fig. 3.2 left part. This option causes the smallest reduction in the data set.
- Mixed set k-fold dead zone cross-validation, where training and test sets are mixed, and pruning may occur in both sets until the dead zone condition holds. Large radius values consume all the data quickly.
- Separated set k-fold dead zone cross-validation, where training and test sets are separated by a gap of radius r. This arrangement uses data more economically, but can be dangerous, since each three data partitions (train-validate-test) should come from as similar a distribution as possible.

Variogram is a well-known spatial dependency test measuring the variance of a variable $f(p)$ over the relative location Δp. Variogram curve starts from value zero and settles to the average variance of the data after a while, thus indicating the spatial correlation zone. The dead zone arrangement gives a chance to a test similar to experimental variogram, where values at roughly similar distance from target points are used to predict the feature value $f(p)$ at points p. This study has been limited to the basic variant of the dead zone where there is only one limitation to the relative locations: $\Delta p > r$.

3.2 Computational tasks

Here the goals set in Ch. 1 are cast to computational terms. The central goals were these:

- To generalise the independent feature Y to areas with no Y measurements, if possible.
- To indicate the possible scope of successful prediction (dead zone testing). This was done with the Pomokaira and Pieksämäki cases.
- To approximate the required density of additional Y measurements at the generalisation area in cases where the generalisation cannot otherwise be performed. Additional samples Y are either in a regular grid or aligned along an access trajectory, see Fig. 2.2. Experimentations with this test will not be included in this project report.
- To recognise environment categories which reach the generalisation level required for practical purposes. This problem is complex and of scientific value but the study had to be excluded from the project due to scheduling problems.
- To find whether additional sets of remote data are needed in order to improve prediction performance. The additional information can be public or private.

There are other goals, which were not approached computationally, but which are the logical next step. One of these goals is to find a widely available set of features which enable the following tasks:

- Finding applicable sub-problems. These include e.g. special environment classes which are both useful and have high prediction performance.
- Dynamical relation between the observed and forecasted weather and terrain trafficability or load bearing capacity of soil. The soil water budget model is the central tool to establish this dynamical relation.

3.3 Machine Learning methods

The methods used in this project are listed here. Some of the properties are mentioned briefly. See (Murphy 2012, p. 14, p. 225, p. 564) and Blaschke (2010) for more information the methods. The literature references are i the order of the occurrence of each method:

- k-NN: simple, intuitive and robust method which uses the average of *k* nearest known samples as the prediction. k-NN loses its applicability in higher dimensional problems, such as the one in this project. The dead zone variant is easy to construct.
- Ridge regression: the sum of total error and total loss terms, both quadratic, is minimised. The dead zone variant uses reduced training sets.
- Multi-layer perceptron (MLP): attempts to minimise target and prediction error through adaption of its inner state (hidden layer weights) accordingly. It can solve both classification (Parkano soil types) and regression problems (Pomokoira water conductivity). The dead zone variant can be arranged in both ways: either by forcing some hidden layer weights to zero or by limiting the input set.
- Object-Based Image Analysis (OBIA) is a set of pattern recognition tools to partition of spatial raster data into meaningful homogenous image-objects, i.e. segments, and to utilise their spectral, shape, and neighbourhood characteristics in order to generate new geographic information. It was used to delineate geomorphological units at the Kemijärvi study area.

4 CASE STUDIES AND RESULTS

The case studies can be divided into four separate categories:

1. Trafficability: there are various case studies on aspects of trafficability on Sec. 4.1. Since there are many cases and diverse approaches, a summary is provided in Sec. 4.1.7. Sec. 4.4. presents FMI hydrological prediction study, which can be a part of the future trafficability forecasting.
2. Remote sensing based monitoring: flood detection in Sec. 4.2, profiling the forest drainage system condition in Sec. 4.3 and storm damage detection in Sec. 4.5
3. Arctic infrastructure: detecting mass flow deposits in Sec. 4.6 and visualising gravel deposits suitable for construction in Sec. 5.2.4
4. Data visualisation: see Sec. 5.2.4.

The sites used were Evo in Southern Finland, Parkano and Pieksämäki, both in Central Finland, and Pomokaira, Kolari and Kittilä in Northern Finland. Detailed information regarding data gathering, theory, test arrangements etc. can be found in the Appendix.

Results and application potential are briefly presented at the end of each case section. Outlines of commercialisation cases are in 4.7. Potential improvements and possible new data sources concerning the set of cases is in Ch. 6.

4.1 Terrain and forest road trafficability

Terrain trafficability in forested areas is currently one of the most important issues in boreal timber harvesting. Trafficability can be hindered due to various reasons, for example steep slopes or surface irregularities. However, most terrain trafficability problems experienced during practical timber harvesting in Finland relate to inadequate soil bearing capacity. Fine grained and organic soils

are particularly sensitive to weather conditions, whereas coarse grained soils are more tolerant. Conducting harvesting operations during periods of sufficient bearing capacity is crucial. Improperly timed operations can potentially cause serious economic and ecological damage. Exceeding soil bearing limitations may also cause tree damage, mostly to roots, but occasionally to trunks as well. This type of damage can lead to fungal infection which eventually causes colour changes in the timber and in the worst case scenario, the decay of the tree. In addition, the water and nutrition conditions of the forest soil can also change as a result of soil settling.

The costs incurred due to challenging trafficability could be reduced significantly with additional information on terrain and soil conditions, especially soil bearing capacity. In addition, forest operations could be scheduled according to periods of adequate bearing capacity or routed to avoid sections of poor bearing capacity, thus minimising damage and maximising harvesting efficiency.

Soil bearing capacity is dependent upon on:
1. Topsoil type
2. Vegetation, especially the strength of the root layer
3. Frost depth
4. The state of the soil water budget

Trafficability also has other factors which should be taken into consideration, including tree root density and reach, stoniness, vegetation thickness, amount of branches scattered on site, topological inclination and various soil layer properties (order, thickness and type of layers). Because of this multiplicity, the term trafficability is a generic label used in this document to bind various numerical aspects together. The final trafficability model may not have just a single index but several. Some insight to this topic is provided by e.g. Pennanen (2003) and Quinn (1991).

4.1.1 Soil type prediction

Soil type is a necessary factor for prediction of the reference load-bearing capacity of a location. Both the soil type and the closely related hydraulic conductivity are useful in estimating the water budget model. The water budget model couples the meteorological dynamical input and the geographical environment then predicts the future load bearing capacity. The test data was gathered from Parkano. The study has been published in (Pohjankukka & Nevalainen (2014).

Parkano site has been fully covered by the aerial gamma ray scan of potassium decay. The concentration of radioactive elements in soil is determined by the source rock and controlled by soil

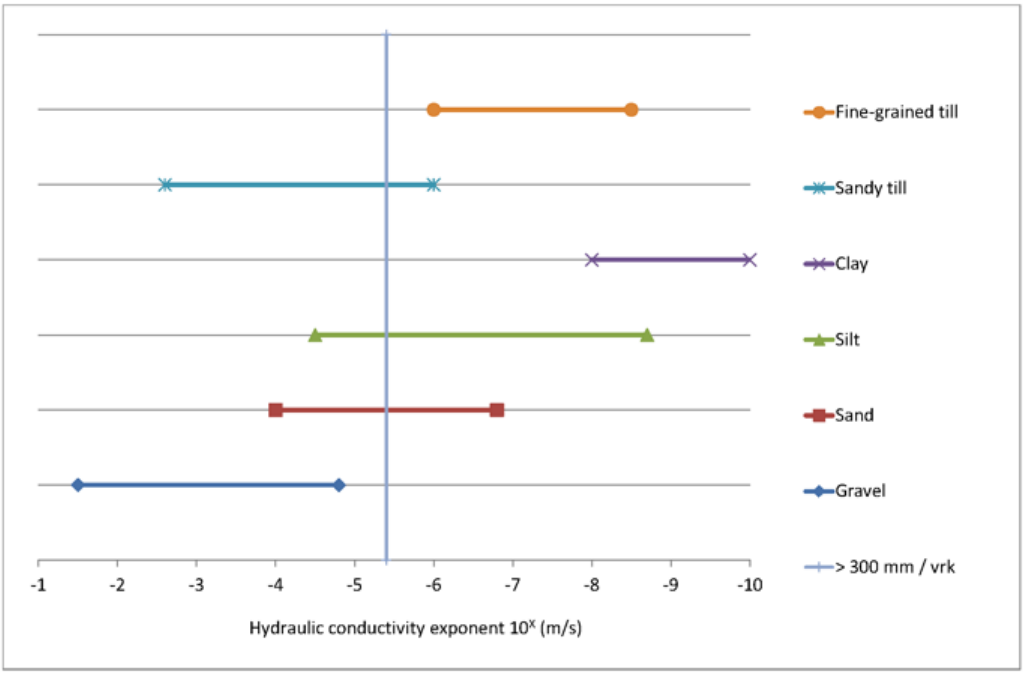

Fig. 4.1. Hydraulic conductivity exponent (directly related to hydraulic conductivity) for various soil types in Finland. Figure by Pekka Hänninen, GTK.

formation and glacial processes. The maximum concentrations are associated with soils developed from felsic rock and clay (Kogan 1971). The soil volumetric water content is the major factor which affects to the intensity variations of gamma radiation in drift materials (Grasty 1997). In Finland, a major part (75%) of the parent materials are tills which have a high spatial variability. The amount of fine-grained matrix is an important factor in determining soil geotechnical properties and it has been proved (Hyvönen et al. 2003) that high water and fine fraction content have low gamma values due to strong attenuation and thus airborne gamma-ray data has potential to aid in the indirect evaluation of the soil material based on their fine fraction content.

The input data was regularly spaced with the raster pixel size of 20 m. Each soil type was given a range of hydraulic conductivity exponent , which indirectly indicates the vertical flow speed of water in different soil types. Hydraulic conductivity exponent ranges for each soil type is illustrated in figure 4.1. The relation between and soil types, albeit not a clear one, provides the opportunity for two formulations: regression to predict and classification to predict the soil type. The soil type 'sandy till' was the easiest to predict in our analysis. This is presumably explained by the high frequency of sandy till (34%) data points when compared with the other soil type data points, causing the training data to have a sampling bias towards these majority soil types.

The input data in this case consisted of the following variables:
- Airborne gamma-ray data, potassium channel
- Topographical height data
- MS-NFI data

The prediction target variables consisted of the following variables:
- Hydraulic conductivity exponent (regression problem)
- Soil type (6 different classes, classification problem)

The methods used in this analysis were: ridge regression, k-nearest neighbor and multinomial logistic regression.

Results: For the regression problem () the best results gave a C-index value of 0.67 and in the classification problems the best results gave 44,5% and 50,5% classification accuracies for multinomial logistic regression and k-nearest neighbor methods respectively. This means prediction (regression,) is possible, but the application should be such that the local distribution of the resulting predictions plays a role. Route planning is one such application, since the relatively modest point-wise prediction performance cumulates to a stronger prediction over longer distances. Route planning however, requires a set of more evolved features, which together define a general trafficability index. These new features may be acquired with the use of new remote data, although this is not within the scope of this project.

Direct soil type prediction (classification) did not perform well. Since the soil type is constant over rather large areas on average, an indirect scheme might be possible:
1. Predict the hydraulic conductivity exponent by ridge regression
2. Use a voting arrangement to find areas potentially containing the same soil type

This two-step scheme will be tested in the future.

Potential applications: It is necessary to input soil type data into the soil mechanics model and water budget model general trafficability prediction. Existing soil class data is based on a set of sparse observation samples and is not very accurate, therefore it may need to be substituted with a predicted quantity. Important distinction between peatbog (see Valjus et.al. (2011), Virtanen (2003)) and other soil types have to be addressed by multisource campaigns.

4.1.2 Vertical hydraulic conductivity related to soil particle size distribution

Gamma-ray data completed with electromagnetic apparent resistivity data is a good combination for studies of near-surface fine fracture geological deposits such as clay and silt (Valjus et al. 2011). The study has been published in (Pohjankukka et al. 2014).

This case had 1788 field measured, rather randomly placed, soil texture analyses. One can derive the hydraulic conductivity exponent indirectly from soil particle size distibution. The rest of the data is grid data and the grid size varies depending on the feature in between 10 m to 50 m, see App.

A1. The site was located in Pomokaira (north of Sodankylä). The goal of the analysis was to predict the hydraulic conductivity exponent mentioned in subsection 4.1.1

The predictor variables in this case consisted of the following:

- Airborne gamma-ray, potassium window data
- Airborne electromagnetic data, apparent resistivity
- Peatland area data
- Topographical height data
- MS-NFI data

In addition to the predictor variables mentioned above we derived numerous features variables to be used in the prediction. These feature variables included:

- Local Binary Pattern texture feature data (Gamma-ray)
- Gabor-filter texture feature data (Gamma-ray)
- Statistical windowed 3x3 mean and standard deviation (Gamma-ray)
- Height features derived from topographical height data (e.g. flow accumulation)

The target variable for prediction was:

- hydraulic conductivity exponent

The analysis was implemented by using the ridge regression and k-nearest neighbor (k-NN) methods. We ran the analysis using three different predictor variable combinations:

- Spatial coordinates
- Input data + features
- Input data + features + spatial coordinates

The spatial coordinates refer to the actual geographical location of the sample data points in ETRS-TM35FIN coordinates. The reason spatial coordinates were included in the analysis was because of an interest in studying the amount of spatial autocorrelation between data points and its affect on the prediction results. In figure 4.2 one can see the 1788 data points plotted in Pomokaira area according to their geographical location as well as the illustration of the dead zone cross-validation method.

Another specialty in this analysis was the dead zone cross-validation (see chapter 3.1.4). It is a variation of the usual Leave-One-Out (LOO) cross-validation. The normal version predicts each point using all the others, whereas the dead zone validation uses all points which are further than a given range away from the point to be predicted. This scheme forces the prediction to only use information which is further away from the prediction point. The dead zone variation can be implemented with LOO and k-fold cross validation, as well as with k-NN method and ridge regression, although with small data sets like this one, the LOO is the more beneficial choice, because it gives the best estimation of the generalisation capability of the prediction model.

Results: The hydraulic conductivity exponent was predictable until appr. 70 m from the nearest measurement using mean absolute percentage error (MAPE). This is a rather weak result, which could require total generalisation i.e. the prediction performance independent of the dead zone

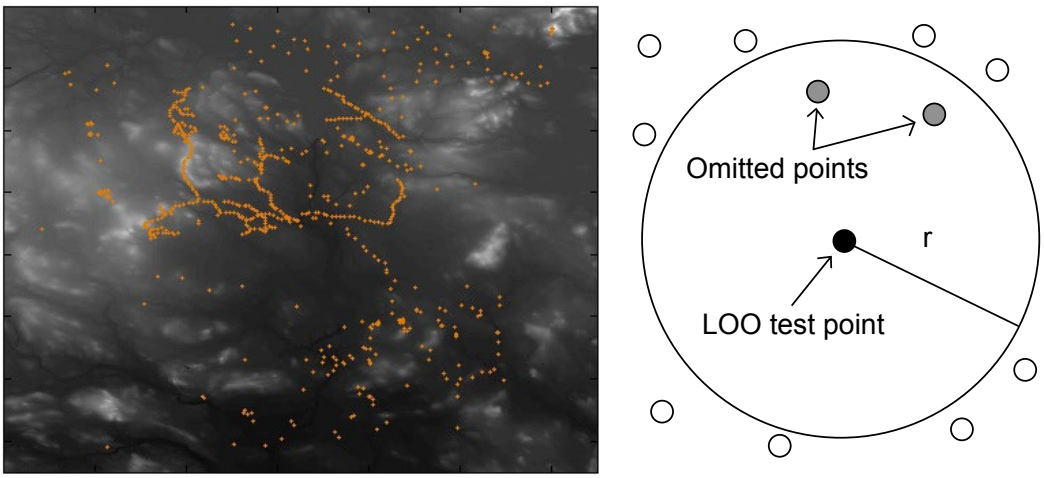

Fig. 4.2. Pomokaira 1788 measurement locations and the dead zone LOO cross-validation method. Figure by J. Pohjankukka, UTU.

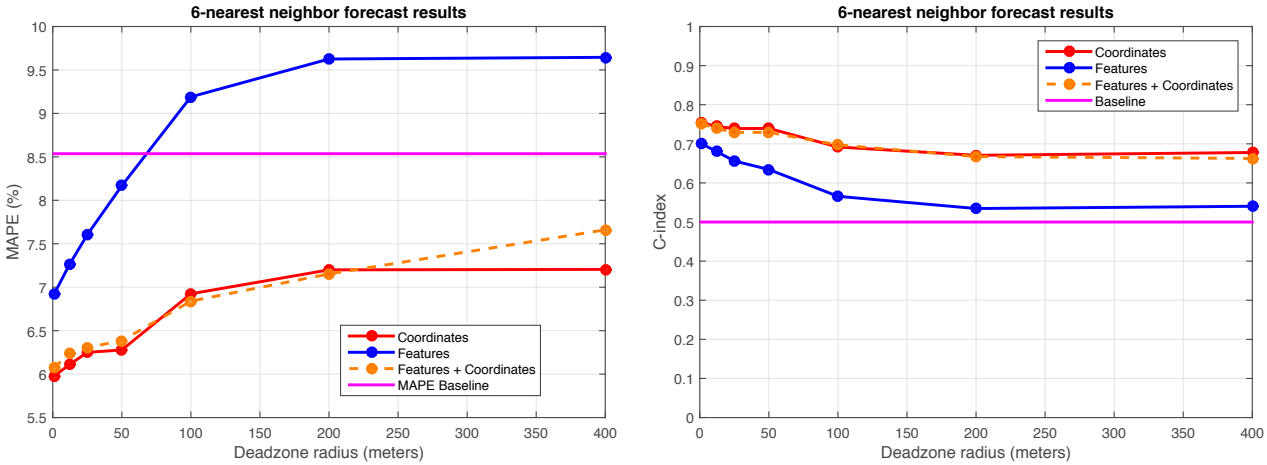

Fig. 4.3. Mean average percentage error and concordance index using LOO and the 6-nearest neighbours. Figure by J. Pohjankukka, UTU.

radius. The results of 6-nearest neighbours are illustrated in figure 4.3. The concordance index (C-index) shows that prediction deteriorates to near random levels within a distance of 200 m. There is a possibility that the prediction range could be increased by modifying the derived texture features and adding a few new primary features e.g. visible and infrared satellite images. Another important observation to make is that the prediction results are better when using the spatial coordinates as predictor variables. Logically this means that there is a large spatial autocorrelation between close by data points.

Potential applications: Predicting the hydraulic conductivity exponent per se has no direct applications. Instead, hydraulic conductivity is one of the key elements of general trafficability, since it is a central parameter in the water budget model.

4.1.3 Load carrying capacity of forest roads

Radar Surface Arrival Detection (RSAD) is a technique to determine ground surface dielectric permittivity by measuring the mode of the travelling time of a ground penetrating radar signal by ensuring the antennae transmission and reception distances remain constant. The signal is related to the average soil density at the top layer of appr. 60 cm thickness. The measured value can be converted into vertical water speed using Saubrei's equation (see Vukovic & Soro (1992)), which relates the soil granularity profile and hydraulic conductivity. Pomokaira case 1 in Sec. 4.1.2 estimated the same attribute more indirectly. The soil porosity (and also the relative air volume) is assumed to be 35%. This is a safe assumption made by GTK experts. Based on this assumption one gets the following three categories, again judged by the geology experts, see Hänninen (1997 and Sutinen (1992b):

1. signal dilatation indicated by the dielectricpermittivity (e_r) : e_r <14 → terrain has good load bearing capacity
2. 14 ≤ e_r ≤ 25 → varying load bearing capacity, moderate load bearing capacity
3. 25 < e_r → no load bearing capacity, not suitable for road construction

The advantage of the ground based radar is that the measurements are related to the corresponding aerial measurements, but the data set of approx. 1900 points is much more accurate spatially. The aerial data has approx. 150 m sampling distance, thus it will not be as sensitive to local variations. The final goal is to base the load carrying estimations on aerial and satellite data, and a necessary initial step is to have a test site with ground based data.

Results: This study will not be concluded during the project, although the data set looks promising. It reflects the same textual pattern of the top and ground soil type. Earlier results are available at Sutinen (1992a).

Potential applications: This case is close to the hydraulic conductivity case in Sec. 4.1.2, but the measurement is more cost-effective, faster and of higher density, so it can produce necessary information about the local variance. This quantity can

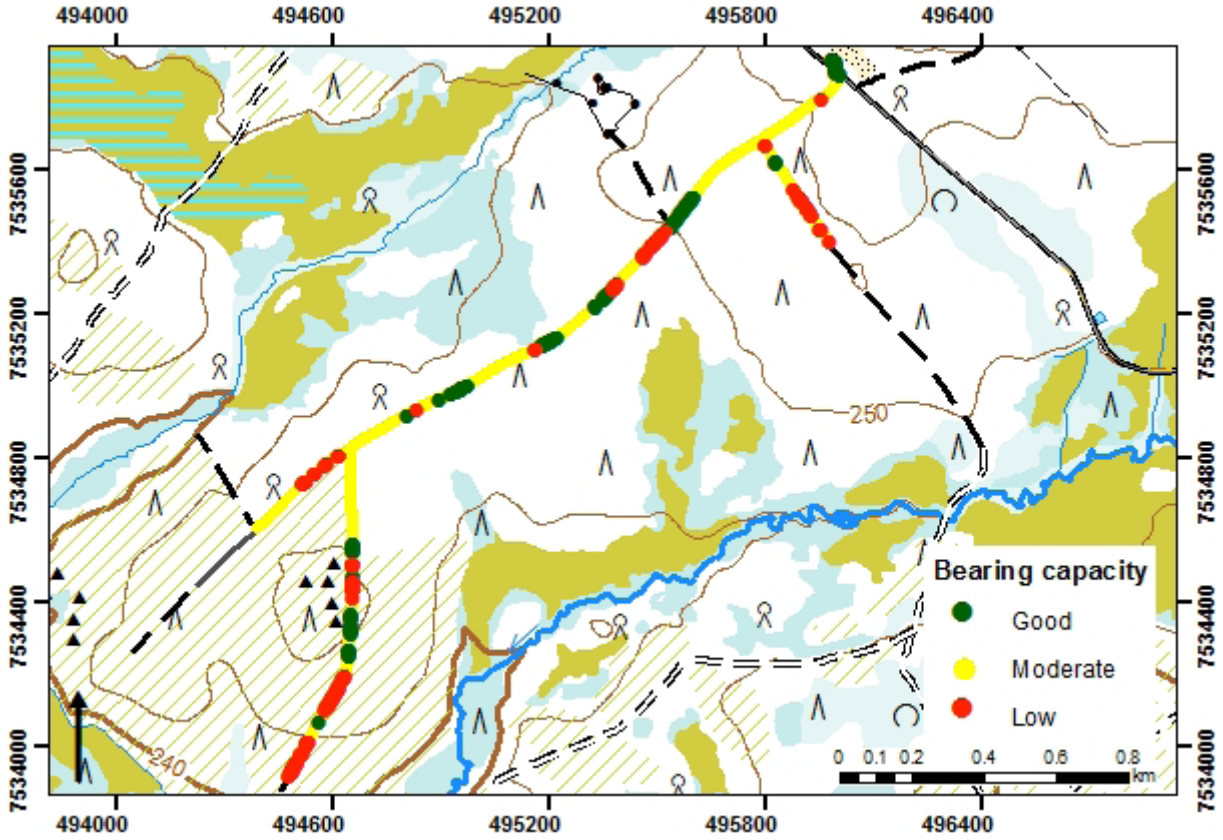

Fig. 4.4. Load bearing capacity category distributions on a road in Pomokaira. Figure by Pekka Hänninen, GTK. Base maps from the National Land Survey of Finland Topographic Database 03/2013 © NLS and HALTIK.

be converted to hydraulic conductivity and it is more directly applicable to the road transportation condition. This quantity, either alone or combined with the hydraulic conductivity exponent derived from public soil data, may improve the predictability. Therefore it may be used for both the water budget and soil mechanics models.

4.1.4 Soil damage prediction, case 1

Approximately 36 km of strip roads were walked through and visually assessed into damage classes by a forest operations expert. This data was kindly provided to us by Metsäteho Oy in 2013. The soil damage data was classified into three (see Table 4.1) main ordinal classes based on the rut depth caused by forest harvesting machinery. The original dataset required preprocessing since the field recorded GPS-tracks included locational errors (zig-zag -motion). After the data was preprocessed to produce smooth line form, we converted strip road lines into points and extracted values from selected features, e.g. MS-NFI and topographic variables. In addition to soil damage classification, the data points were classified based on the strip road type and the harvesting operation type. The three different data point classifications are highlighted in Table 4.1.

Table 4.1. Classification of the soil damage data points. Table by J. Pohjankukka, UTU.

Damage class	• No damage • Slight soil damage • Soil damage **
Strip road type	• Primary (high traffic intensity) • Secondary (low traffic intensity)
Operation type	• Thinning • Clear cutting

**=Includes strip road sections covered by brash mat, originally classified as "potential damage", since without brash mat they likely would have been damaged.

In summary, a data point had three potential types of classification: damage class, strip road type and harvesting operation type. Analyses were conducted for each of these cases separately to keep the data sets consistent. In this section, the results of the analysis for secondary strip roads comprising 76% of the whole data set are reported.

The aim of the analysis was to a study the use of public data in predicting soil damage caused by

harvesting machinery. Trafficability forecasting for forest operations requires the prediction of local soil layer properties as well as both local and wider area topography. Various geomorphometric variables which relate to soil (moisture) properties were calculated in IS from NLS ALS Data, e.g.:

1. Flow accumulation area
2. Local slope
3. Topographic Wetness Index
4. Potential incoming radiation

In addition we used:

1. Multi-Source National Forest Inventory data, 43 variables (publicly open)
2. Soil type data (publicly open)
3. Peatland mask data (publicly open, created from NLS topographic database by Metla)
4. Gamma-ray spectroscopy data (public, GTK)

The target variable for prediction was:

- Soil damage (3 classes)

Numerous statistical and textural features were derived from gamma-ray data such as windowed mean, windowed standard deviation, Gabor-filter features and Local Binary Pattern features. See full list of variables at appendix A1. Vertical distance to drainage network (or depth to water) has been successfully tested in Canada for the purpose of land bearing capacity modelling (Murphy et al. 2009). However the open data of NLS topographic database depicting the drainage network is not comprehensive, and therefore we could not use this approach (see sect 4.3, Fig. 4.14). Many of the ditches are missing and there are locational errors, due the fact that these product has been done, to our knowledge, mainly based on aerial image interpretation and therefore ditches under dense canopies are hard to detect. This observation was one of the primary reason to do study regarding drainage network detection (see sect, 4.3).

The Machine Learning methods used were the k-nearest neighbour algorithm and ridge regression (a.k.a Tikhonov regularisation method). The optimal model parameters were estimated by using the Leave-One-Out cross validation method. In addition, the predictive performance of the model as a function of the spatial distance between the training data and test data was estimated. This was done in order to infer how far away from the current location we can obtain reliable predictions. This is particularly useful for example in route planning for the harvesting machinery. The typical train-validate-test sampling arrangement was modified by using the dead zone approach presented briefly in Sec. 4.1.2.

Two different performance indicators were applied in evaluating the model:

- Concordance index (CI)
- Percentage of successful soil damage classifications

The concordance index measures how well the forecast model captures the relative order of data pairs. In other words, the forecast model got the prediction right if the relative ordering of predictions is the same as it was with the values in the reference data . If the relative orderings are correct most of the time, the forecast model successfully captures the pattern from the data. CI is a real number in the range where the value 0.5 means that the model captures only noise from the data, 1 indicates a perfect match and 0 indicates that model describes the pattern perfectly but in reverse order. A model with a CI value approaching 0 or 1 has good prediction capability. Percentage of successful soil damage classifications indicates the frequency of successful classifications by the model.

Results: The results for soil damage prediction and classification as functions of the deadzone radius are depicted in Fig. 4.5.a, i.e. the spatial distance of test point from the closest training points. In figure 4.5.b we have illustrated the results for the same analysis, but with added weather data information (mean temperature and rainfall of past month) into the prediction model.

The results indicate a satisfactory predictions up to 20 m deadzone radius. This means that forecasts can be made with sufficient accuracy merely up to 20 m range from current location of the harvester with the data set used. After 20 m the prediction performance drops dramatically and a random yes/no guess produces better results. We notice also that the added weather data in the model improves the prediction accuracy. CI value stays above the baseline until 200 m in the ridge regression case, whereas without weather data the CI stayed above baseline up to 75 m. Classification results improve about 30% with weather data, but acceptable prediction rate (above 50% baseline) reaches only 20 m prediction radius with or without weather

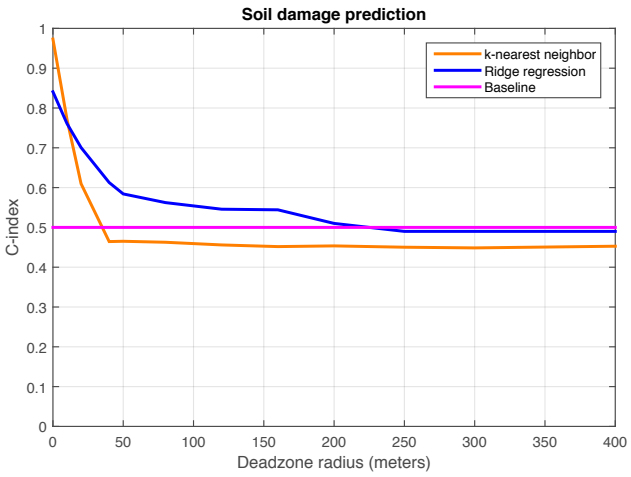

Fig. 4.5.a. Soil damage prediction with k-nearest neighbour (orange) and ridge regression (blue) methods as indicated by concordance index (C-index, also CI) and rate of successful predictions. Figure by J. Pohjankukka, UTU.

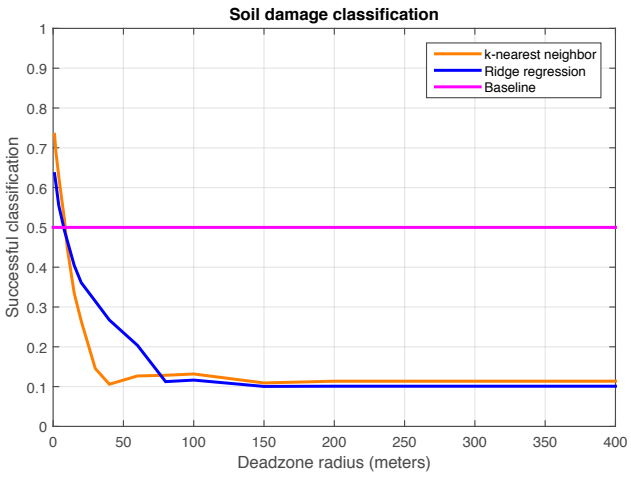

Fig. 4.5.b. The same results as in figure 4.5.a. but with added weather data in the prediction model. Figure by J. Pohjankukka, UTU.

data. In summary, weather data improves the results but not enough to have significant predictive value for practical purposes.

At the final stage approx. 11000 sample points at 2 metre intervals were used. The predictions were divided into separate cases of clear cutting and thinning and further primary and secondary strip roads. The four combinations formed differ slightly and the practical applications presumably can and will use similar classification. Secondary strip road was characterised by a low number of passes leading to better prediction performance. This is because the amount of damage is cumulative, and more passings likely bring more damage, and having the number of actual passings unknown, it is wise to use only the data set with minimum number of passes i.e. secondary strip roads.

We conclude, that a more detailed field reference is needed, i.e. physical measurements and detailed information about forestry machinery movements within the stand, since the accumulation of traversed mass over each location is one of the main variables explaining soil damages (Siren et al. 2013). These can be achieved through online learning based on data accumulated by harvesters or another field study (see sect 6.4.3). Vertical distance to drainage network should also be tested (see sect. 4.3). Further validation with weather data and water budget models should be continued in the future as it is one of the key variables affecting trafficability of fine grained mineral and organic soils. New test cases from varying environments are needed in the future.

Potential applications: With additional information timber procurement can be planned more efficiently, thus lowering cost and minimising ecological issues, such as damage to tree roots and undesirable soil deformation. In the future the inclusion of dynamic water budget models will improve the models further, providing more explanatory power and a better generalisation capability (spatial independence). In practical terms this means good performance with a wider dead zone radius. Moreover, these models can be updated and calibrated based on harvester CAN-bus data (see Ch. 6).

4.1.5 Soil damage prediction, case 2

Instead of relying solely on damage class data we decided to make supplementary soil strength measurements describing soil load bearing capacity. We measured a total of 50 locations on two different stands in Kumpunen, Pieksämäki, Finland (N 6921354 E 501297). The study was conducted during a commercial thinning operation on two different stands. The selection of plot locations was based on expert judgment to cover the gradient between dry mineral soil with high bearing capacity and wet organic soil with low bearing capacity. Depth of the organic soil varied from 0 to almost 90 cm. We measured the soil penetration resistance using a penetrometer (Pohja 2000, Muro and O'Brien 2004) at five different locations around and between the tracks to avoid the random effect caused by e.g. hitting a tree root in a single

measurement. Also shear modulus and stoniness was measured similarly at five different locations.

Penetrometer hits to stones, resulting into an interruption of the measurement, are usually given the last measured resistance value. This normally equals or approaches to the maximum capacity of the measuring device, yet it has nothing to do with the encountered penetration resistance of the cone-stone combination. As a first approach towards a more realistic imitation of the real phenomena, penetration resistance was determined as the resisting force met by a typical stone penetrating into the soil similar to that immediately above the stone divided by the cone basal area. The cross sectional area of the typical stone was given the value 0,00281 m^2.

Stoniness was estimated by steel-rod sounding (Tamminen 1991). The rod was pushed into the soil where the penetration depth and stone hits were recorded. Shear modulus measurements were performed with a spiked shear vane (Ala-Ilomäki 2013). The accumulation of logging residue significantly hindered measurement, meaning it wasn't always possible to place measurements systematically.

The depth of the wheel rut was measured using an inverted U-shaped frame with its feet resting on the undeformed soil surface outside the wheel rut, which formed the reference level. Individual observations were averaged to plot level. The first measurements were taken after the harvester passed, and the rut formation was measured again after each pass of the forwarder collecting the timber from the cutting area. The extraction road was cleared of logging residue after the harvester pass in order to observe the effect of soil properties on forwarder rut formation without the reinforcing effect of brash (Siren et al. 2013). The accumulated mass traversed over each measuring location was defined as the sum of net vehicle mass plus the mass of load for all the passes (Siren et al. 2013).

The aim of this analysis was to study the capability of soil bearing classification to predict rut depth after harvesting machinery. Amongst the measured variables, soil penetration resistance was best describing soil bearing capacity in the varying soil conditions. Soil was classified as bearing or nonbearing based on the magnitude of penetration resistance with the decision boundary being 5000 kPa. This means that soils which have a penetration resistance of less than 5000 kPa are classified as nonbearing regarding harvesting operation purposes, whereas soils with penetration resistance equal to or more than 5000 kPa were classified bearing. The data consisted of a total of 50 data points from two separate harvesting sites.

Three different machine learning approaches were tested with both linear and nonlinear methods. The methods used were: k-nearest neighbour, ridge regression and multilayer perceptron model. Validation and model selections were implemented using the Leave-One-Out deadzone method as used in chapter 4.1.4.

The predictor variables, i.e. input data for this task consisted of the following data sets:

- MS-NFI data (remote sensing)
- ALS data (remote sensing)
- Stoniness data (field measurement)
- Peat data (field measurement)

The two prediction target variables were:

- Penetration resistance, regression
- Soil bearing capacity class (soil penetration resistance over or under 5 mPa)

The same performance indicators as in chapter 4.1.4, namely concordance index and successful classification rate, were used with the 50 field measurement points. The results of this analysis can be seen in figure 4.6.

Results: The prediction capability is definitely better than it is in the subsection 4.1.4 case. The successful classification rate stays above 50% of the baseline value up to 110m and the CI value up to 120 m. Geographically close predictions have a very high classification rate of 85% up to 20m.

Deadzone with a 100 m radius tends to be the limit point for predictions degrading to a level below a random yes/no decision. We can also see from the graphs that ridge regression and multilayer perceptron (MLP, see Sec. 3.3) tend to have similar performance rates with ridge regression showing slightly more stable results due to the linear nature of the model. k-nearest neighbour on the other hand shows the lowest prediction performance and is very sensitive to the increment of deadzone radius.

As a conclusion the predictions remain very good up to 20 m, good up to 50 m and above the baseline until the 100 m boundary. It was evident that the used data set was much more reliable and

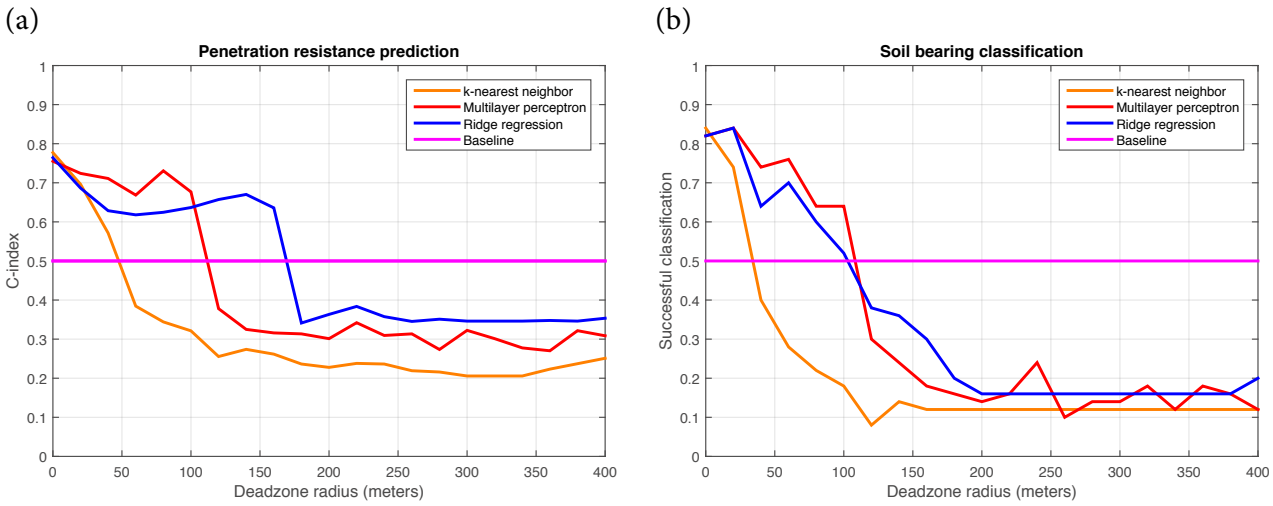

Fig. 4.6. Real value prediction capability of penetration resistance (regression) measured with C-index as a function of deadzone radius (a) and soil bearing classification capability with the success rate (%) as a function of deadzone (b). Figure by J. Pohjankukka, UTU.

contained less noise than the data used in the 4.1.4 case. The data set consisted of real measurements and recorded traffic intensity instead of visual estimations and assumptions. This points out the necessity of accurate and real-time measurements in order to produce applicable forecast models for harvesting operations. If the data quality is not sufficient, the probability for prediction capability with beneficial performance rapidly deteriorates.

Potential applications: It is likely that the soil penetration resistance and soil type classification can be predicted generally, i.e. given more test sites from various environments, a general predictor covering meaningful ranges and spanning the whole of Finland can be formed. We have to test the prediction by using e.g. separate models of different environments (multi-model regression, where an initial decision tree indicates which regression model should be used). There is an important connection to online learning based on the data gathered by forest harvesting machinery. Even with the limited data used here the results were surprisingly good, especially compared to case 1 in Sec. 4.1.4. It seems the high-quality input data significantly improved the models. In the future the data can be derived for example from the CAN-bus of harvesting machinery (transmission power expenditure), supplemented by rut depth and accurate micro-positioning using mounted LiDAR. Even only one attribute from this set would be a useful online learning data source.

4.1.6 Forwarder route selection

In section 4.1.4 the aim was to minimise the soil rutting caused by harvesting operations. The analysis was implemented by predicting individual soil points one at a time (Leave-One-Out cross-validation). The results indicated rather low prediction capability with the provided data set. Here, another approach is being tested. The idea is to cumulate the predictions of individual soil points to evaluate the prediction performance on alternative paths instead of individual soil points. The baseline case assumes no specific information on the paths and always chooses the shortest path. This is an attempt to set a cost term to route planning, as presented in Väätäinen et. al (2013).

The aim of this study is an attempt to overcome the poor performance of pointwise predictions with a cumulative measure. The test case (the base line) is depicted in Fig. 4.7. A forwarder has a task to move from A to B and since the case has no dense measurements available, the existing data is used for two route alternatives. The length difference distribution of alternatives was controlled with practical limits. The damage cumulation was done by giving weights , to three damage classes. The geometric ratio was given by experts (Metla). The formulation of the damage cumulation law can be later changed to best possible in terms of prediction efficiency. The same predictor and target variables were used in this study as in the case of subsection 4.1.4.

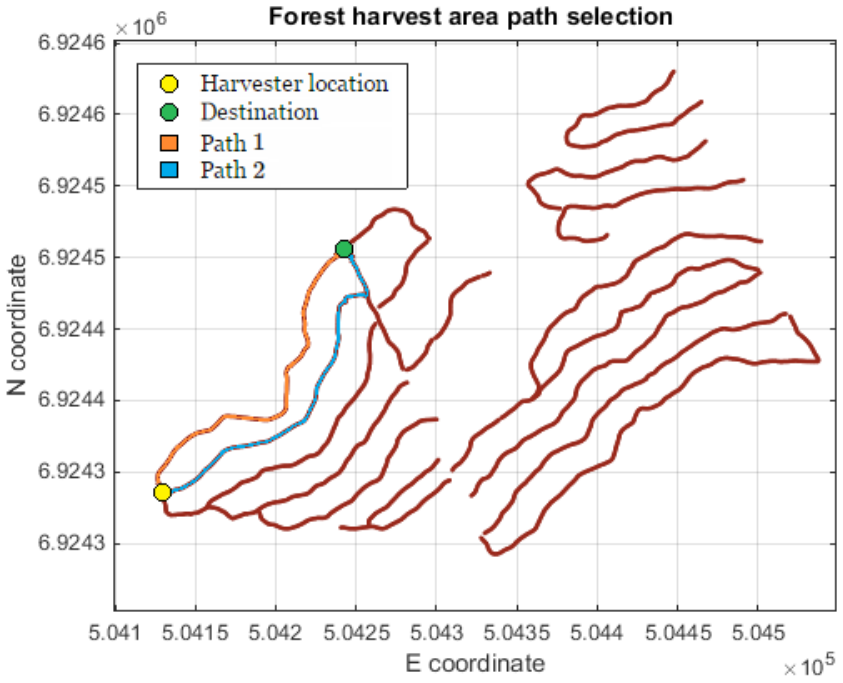

Fig. 4.7. Illustration of forest harvest area with a path selection application. A forwarder has a starting location (yellow point) and a destination (green point). The problem is to select the best path in order to minimise the total damage caused to the soil. Figure by J. Pohjankukka, UTU.

Results: The results indicate a more than 80% success rate when attempting to choose the better of two alternative harvesting paths. In other words even though the prediction model was not able to give acceptable prediction rates for individual pointwise predictions it still succeeded in classifying most of the time correctly the better path of the two, i.e. the route that had the smaller value of total soil damage.

The estimate of visual soil damage indicates physical consequences of harvesting and has correlation to future economic losses - but it is not a measured physical variable. We can conclude that it is better to focus on soil penetration resistance, rut depth and anything which can be directly or indirectly measured by forest harvesting machinery.

Potential applications: This case addresses another cost function ingredient in forest harvesting operations, i.e. the cumulated damage and risk of exceeding the load carrying capacity. There is no need to have a special application for predicting best routes, instead a trafficability application needs the functionality to be able to suggest routes to the forest machine driver.

4.1.7 Trafficability cases, a summary

The prediction results are relatively modest in most cases except with the cumulative prediction path selection problem (Sec. 4.1.4). This is not a case pro visual classification and not a permanent proof against pointwise prediction, since the good result derives from the cumulative nature of the indicator in Sec 4.1.4. The similar improvement would show up in case in Sec. 4.1.6, if a cumulative formulation were to be applied.

There are are three possible roads to pointwise prediction improvement:

- finding new wide area primary features, e.g. various satellite imaging, aerial electromagnetic scans, geomorphological features, etc.
- larger scale texture features and other methods to generate derived features.
- new Machine Learning methods and their variants. One of them is the rejection of the noisy part of the problem. Experiments with Parkano data indicate, that by limiting the teaching and prediction to a practical subset of locations, the performance on the resulting limited area can be improved.

4.2 Flood detection

Since autumn 2012 FMI has been developing methods of flood detection and mapping based on satellite remote sensing. COSMO-SkyMed data was used to create flood maps of the Pohjanmaa, Kittilä, Kolari and Evo regions. COSMO-SkyMed consists of four individual satellites equipped with X-band Synthetic Aperture Radar (SAR) sensors. The major advantage in using SAR instead of optical sensors is the ability to penetrate clouds, and therefore SAR in general is the ideal sensor type for flood mapping in the Finnish climate. The COSMO-SkyMed constellation in particular is suitable for mapping natural hazards and damage because it enables daily imaging of any region on the earth. Moreover, FMI is receiving Cosmo data straight through a receiving station located in Sodankylä. COSMO-SkyMed offers several imaging modes, where the spatial resolution ranges between 0.5 m and 100 m. In the future the COSMO-SkyMed data could be replaced with publicly open SAR products.

A semi-automatic algorithm was created to enable fast processing of the satellite data in order to provide near real-time flood maps. In the beginning, the focus was only on flood detection in open, treeless areas. The methods were tested for the first time in the Pohjanmaa region, during the autumn floods of October 2012. The floods were successfully recognised and the total processing time, which included satellite image downlink (~2 h), analysis (2-4 h), visualisation (1 h) and dissemination (1 h) took approximately 7 hours. In April 2013 mapping of the spring floods in Pohjanmaa was carried out operationally, and results were delivered to Finnish Environment Institute (SYKE) approximately 6 hours after the satellite acquisition. This serves as a proof of the feasibility of this technique.

The scattering of the SAR signal when reacting with flooded forests is different than when reacting with open area floods. When hitting water under tree canopy, a large portion of the SAR signal returns back to the sensor due to water-tree trunk reflection, creating strong backscatter in the signal (McDonald et al. 1980, Engheta & Elachi 1982, Richards et al. 1987, Townsend 2002). In open floods there is so called specular reflection creating a weak backscatter signal. In order to study flood mapping in forested areas, SAR data was collected in 2014 during spring floods in the Kittilä, Kolari and Evo regions, from both forested and open areas. On site measurements were conducted by Metla for the purpose of evaluating the satellite interpreted flood areas. Flood detection in different forest densities and heights was investigated using LiDAR-based canopy closure and tree height data.

Results: Floods were well detected in forested areas with normal forest properties. However, the detectability of floods was reduced in forests with lower tree height and sparse forests. It was also noticed that the combination of tall trees and shallow look angle produced a geometrical shift (up to 30 m in Evo) in the satellite interpreted flood areas. Tree height was the forest parameter which had more influence on the flood detection accuracy than CC. In order to improve the results in forests with low dominant height, SAR-detected floods were expanded to cover them by using other spatial data in addition to the SAR images, such as LiDAR based forest maps, high resolution digital elevation models and land cover classification data. According to preliminary results 90 % of the floods were detected in Kittilä, 85 % in Kolari and 60 % in Evo. The weaker results from Evo were due to the geometric shift caused by tall trees and shallow look angle.

Potential applications: The advantage of satellite based flood mapping compared to traditional methods such as aerial photos and ground observations is the ability to map large areas in a fast and cost efficient way. Near real-time flood maps can be incredibly beneficial to emergency personnel operating during a flood event, as well as to landowners and farmers when planning preventative/recovery actions. Flood maps can also be used by insurance companies when handling flood related claims and by other private or public sectors for community, agriculture and forest planning. The progress of the floods could also be predicted more accurately during the spring snow melting period based on flood maps of the previous years.

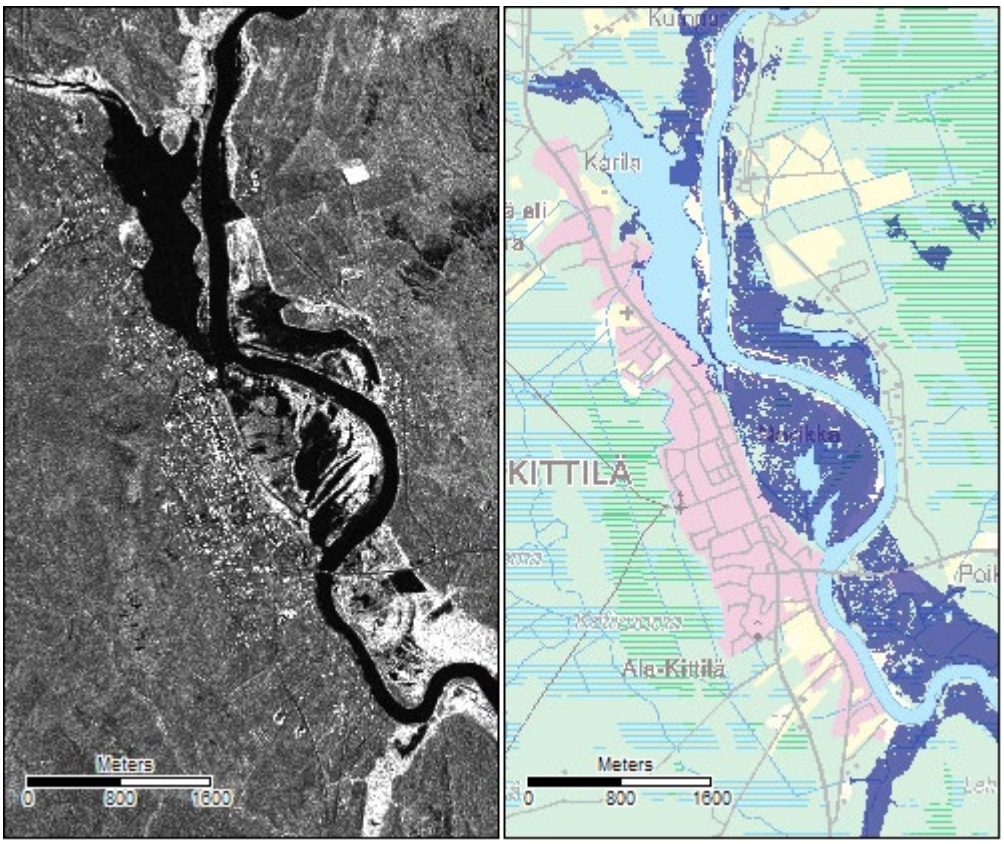

Fig. 4.8. Left: Cosmo Sky-Med SAR intensity image from Kittilä. Open floods, lakes and rivers have low backscatter and appear dark on the SAR image, whereas forest floods create high backscatter and appear white on the SAR image. Right: Detected open and forest flooded areas marked as dark blue from the same area as the SAR image. Figure by Juval Cohen, FMI.
Base maps from the National Land Survey of Finland Topographic Database 03/2013 © NLS and HALTIK.

4.3 Forest drainage network extraction based on airborne laser scanning

Vertical distance to drainage network (or depth to water) has been successfully tested in Canada for the purpose of soil bearing capacity modelling (Murphy et al. 2009). However the open data of NLS topographic database depicting the drainage network is not comprehensive, and therefore we could not use this approach (see sect 4.3, Fig. 4.14). Many of the ditches are missing and there are locational errors. This observation was one of the primary reasons to do a study about drainage network detection based on ALS data and data mining techniques, leading to this developed application. We believe that results from drainage network extraction can be used as an input in future soil bearing capacity modelling or by NLS if a more detailed base map is desired.

The high resolution and accuracy of airborne laser scanning (ALS) makes it an ideal technique to use in fine-detailed morphological mapping. Here we utilised open low-pulse density (0,5 p/m2) NLS ALS data and high-pulse density (10 p/m2) ALS data to investigate the potential of the open NLS ALS product, specifically for the mapping and evaluation of drainage network coverage and condition via remote sensing. The study site was located in Evo, southern Finland, where 62,5 hectares of drained forest land were thoroughly covered on foot in order to record ditch locations and condition using a high-accuracy GPS-receiver.

Digital elevation models (DEMs) from the point cloud data were first created using Lastools and SAGA GIS software. Points with scanning angle larger than 20 degrees were removed. If excluding 20 deg. echoes resulted to large no-data areas a 25 degree threshold value was used instead. The ground elevation was then detected, and points higher than 0.5 m above the detected ground level were removed. The remaining points were gridded to 0.25 m and 0.5 m resolution from the higher resolution and lower resolution laser data respectively. Two different gridding methods were tested, nearest neighbour (NN) and inverse distance

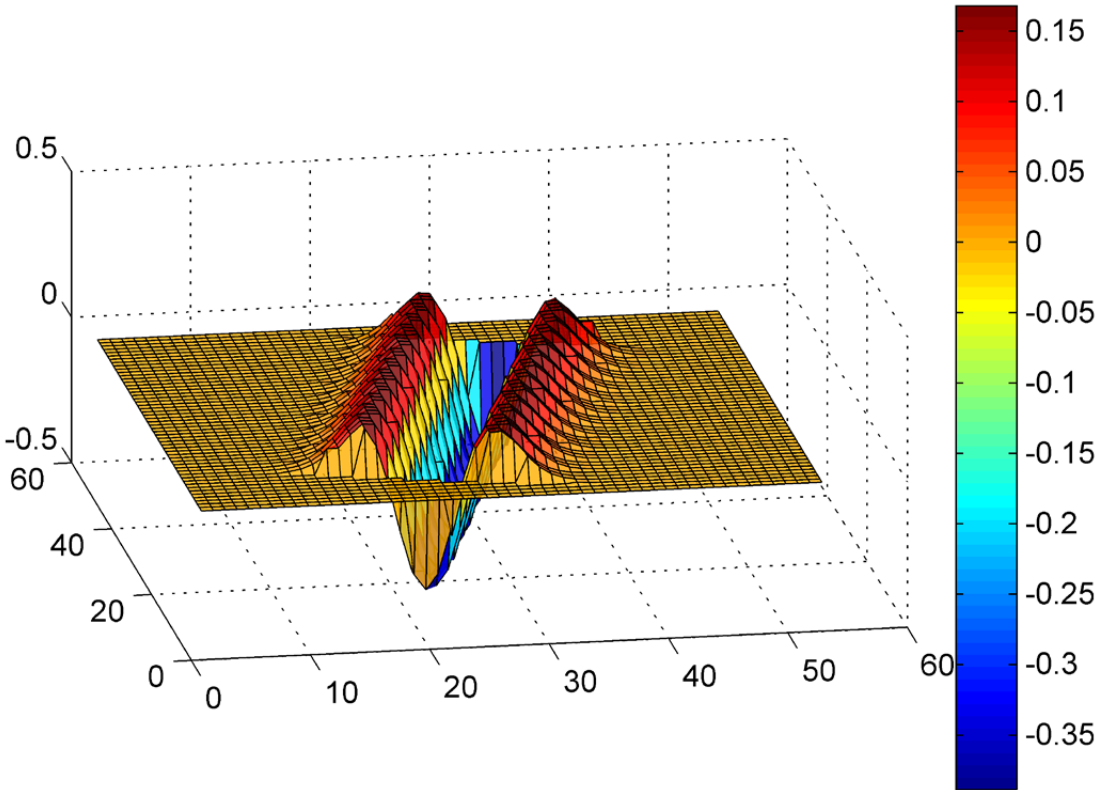

Fig 4.9. The second derivative filter in pixel coordinates. The filter was applied on the DEMs in order to extract the ditches. Figure by Juval Cohen, FMI.

weighted interpolation (IDW). The ditches were detected with a second derivative 2D filter:

$$f(x) = -\frac{1}{\pi\sigma^4}\left[\frac{1}{2} - \frac{x^2}{2\sigma^2}\right] e^{-\frac{x^2}{2\sigma^2}} \qquad (4.1)$$

, which is based on the 3D LoG (Laplace of Gaussian) filter:

$$LoG(x,y) = -\frac{1}{\pi\sigma^4}\left[1 - \frac{x^2+y^2}{2\sigma^2}\right] e^{-\frac{x^2+y^2}{2\sigma^2}} \qquad (4.2)$$

The Gaussian standard deviation σ was set to 0.8, which was the best for detecting typical sized ditches. The filter was rotated to 32 different angles ranging from 0 to 31π/32 (e.g. 0,π/32,2π/32,3π/32 …31π/32) using nearest neighbour interpolation, and applied to the DEMs, giving positive values to ditches and negative values to embankments. Then, for each DEM pixel the highest value among the 32 filtered values was chosen. In order to remove falsely detected ditches due to large ground slopes, the DEMs were filtered with a first derivative filter, and pixels with a slope larger than 5 degrees were masked. After testing different filter sizes, 10 x 10 m filters were chosen, which is equivalent to 21 X 21 pixels for the low-pulse density data and 41 x 41 pixels for the high-pulse density data.

After the DEM is filtered it needs further processing to retrieve the actual drainage network. This is done with threshold values. Firstly, we set a minimum value of 2 to the low resolution 2nd derivative, and 10 for the high resolution 2nd derivative, respectively. In addition, a minimum segment size was set to 20 m² to mask pits and other ditch-like landforms which are not connected to the actual drainage network. Thresholds were selected based on simulations which were run with manually created DEM's. 2nd derivative value is dependent on the filter size used. At the end the drainage network raster was thinned to single cell width in ArcGIS and then converted into polyline (Zhan 1993, ESRI 2014).

Results: The preliminary results of the analysis look very promising, as in the best reference sites almost every ditch was detected (Figs. 4.10, 4.11 & 4.14). Moreover, we can estimate the depth of the ditch based on airborne laser scanning data although there is high uncertainty regarding a single observation (Fig. 4.12). Some of the uncertainty arises from the fact that depth is measured from a single point, while the DEM derived depth is calculated from an area of 100 m² . The depth

calculated from the DEM is generally an underestimated, which is probably caused by the lower accuracy of ALS compared to field observation and the ALS footprint size, which smooths the true terrain surface. Furthermore, the IDW interpolation smooths the surface even more. Despite these uncertainties we are confident that a product of this kind will advance hydrological models, since we get more accurate drainage networks (density, length), and we can estimate the depth of an individual ditch (Figs 4.10-4.14).

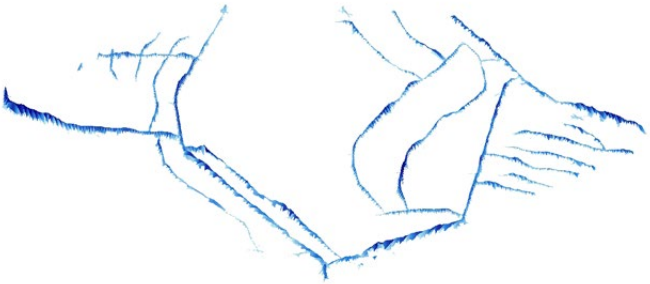

Fig. 4.10. 250 x 250 meter area located in Evo, and its drainage network visualised in 3D with the 2 nd derivative filter values (DEM: low resolution, IDW). Deeper blue and higher Z value indicates a deeper ditch and good draining capacity based on ALS data and LoG-filter analyse. Figure by Henri Riihimäki, Metla.

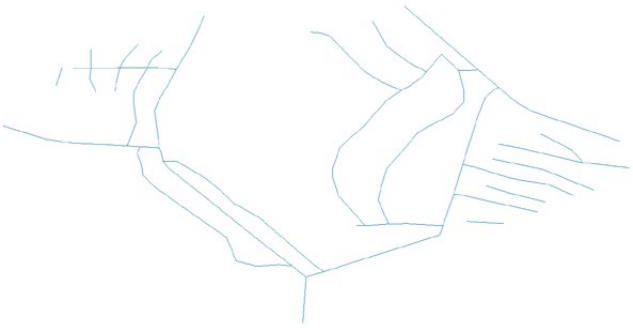

Fig. 4.11. Field recorded reference drainage network from the same 250 x 250 meter area. Most of the ditches are found with the algorithm. Minor gaps are found in some of the ditches. Figure by Henri Riihimäki, Metla.

Figure 4.12 depicts the correlation between the observed and calculated ditch depth (rho = 0.44). The DEM based ditch depth was retrieved by calculating the difference between DEM value and the mean elevation based on low resolution DEM (Gridding method is nearest neighbour, NN). Mean elevation was calculated from 10 x 10 m window where ditches were first masked off. Blueline is the trendline between the two.

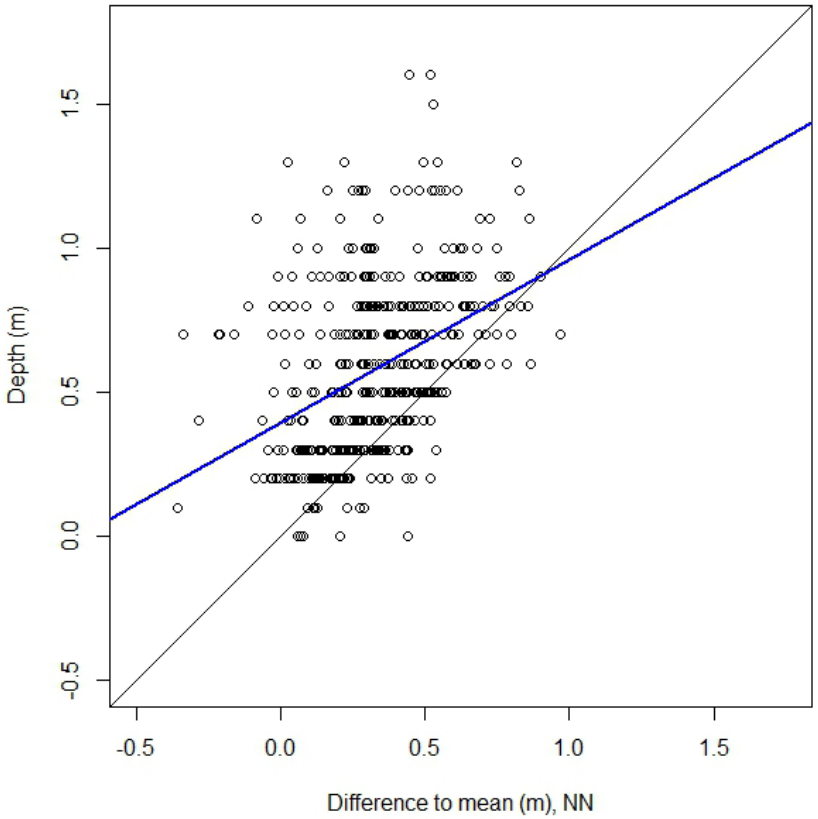

Fig. 4.12. Ditch depth observed on the field (448 locations) compared to the calculated ditch depth from the DEM at the corresponding location. Figure by Henri Riihimäki, Metla.

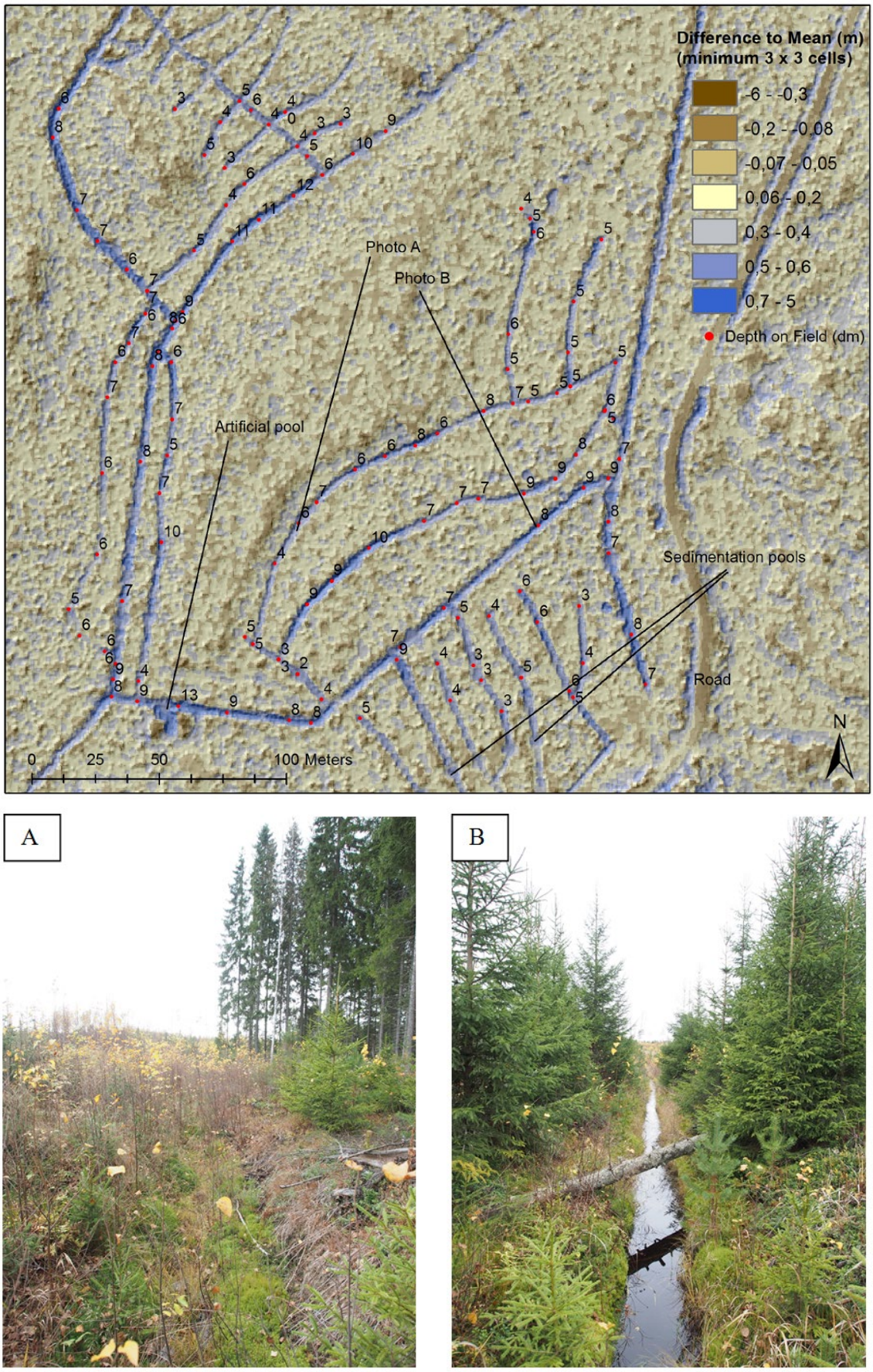

Fig. 4.13. 2D -visualisation of drainage network depth and depth measurements on field (in dm). In photo A the ditch condition is poor and it is shallower than that of B. The A picture is taken towards SSW and B towards SW. Photos: H. Riihimäki, Metla.

The second derivative filter has many variations depending on which kind of longitudinal profile it has. Experiments with different parameterisation and layout of the filter will continue, but in general the combination of the specifically tailored filter and LiDAR derived DEM data seem to be a very good choice. We consider that it is a suitable tool for updating the NLS topographic database, if desired, and in the creation of comprehensive drainage networks for other purposes as well. Some noise and false ditches could be removed by Artificial Intelligence methods, which could automatically correct the drainage network to a technologically likely state before using it in another application.

Potential applications: With this methodology it is possible to create comprehensive descriptions of existing drainage networks. Therefore, this could be used to update the NLS topographic database and base maps, although in a few reference areas the results did not result into a large improvement. The precise location and condition of the drainage network can be utilised in the water budget modelling and it also helps in route planning of harvesting operations. Moreover, Metsäkeskus, Metsähallitus and individual forest owners and companies could benefit from this application, e.g. when planning drainage network maintenance. In addition to ditches, other morphological forms of the network, such as sedimentation pools can be detected with the analysis (Fig. 4.13) The results can be further utilised in hydrology, forest trafficability (see Sec 4.1.), but also in ecological applications, which need information regarding soil wetness, as drainage density and the condition of the ditches both have a significant influence on soil wetness (Päivänen & Hånell 2012).

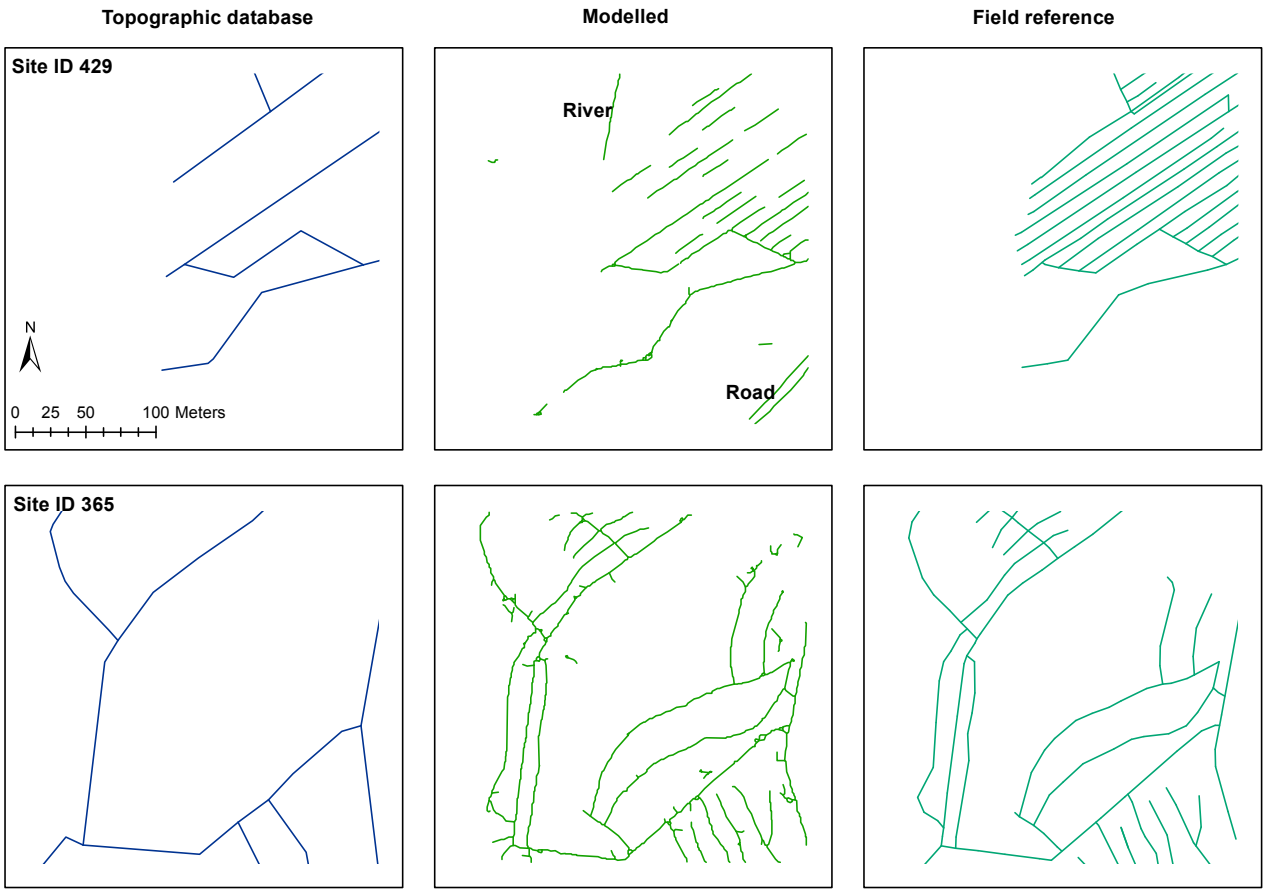

Fig. 4.14. Comparison of the drainage networks of NLS topographic database, modelled and field reference on two selected sites. Grid size is 250 x 250, DEM: Low Resolution, Figure by H. Riihimäki.

4.4 Hydrological Operations and Prediction System, HOPS

A distributed (gridded) Hydrological Prediction and Operations system (HOPS) was developed during the project. With the modelling system it is possible to configure models anywhere within the geographic extents of Finland, using the currently readily available GIS and meteorological forcing data. The modelling system can be used for historic analysis using historical meteorological data as forcing or in forecast mode, where e.g. European Centre for Medium-Range Weather Forecasts (ECMWF) and/or High Resolution Limited Area Model (HIRLAM) meteorological forecast data can be used to drive the modelling system's modules. It is also possible to assimilate earth observation (EO) data on snow water equivalent, soil moisture and soil frost conditions into the modelling system.

The HOPS system is modular and contains the following components:

1. A gridded soil moisture model based on the Sacramento Soil Moisture Accounting Model SAC-SMA, see Burnash (1995).
2. A distributed river routing model accounting for overland-, and channel flow retention and attenuation, based on hydrographic properties.
3. A potential evapotranspiration model based on Hamon's approach in Oudin et. al (2004).
4. An in-house developed temperature index snow model.
5. A soil temperature model based on Rankinen et. al (2004).

The modelling system's forcing data requirements are very low compared to other more complicated modelling systems. The forcing data requirements include daily precipitation sum, daily average temperature and daily sunshine hours sum. The parameters for the SAC-SMA model (the core of the modelling system) can be derived from a semi-physical a-priori parameterisation schema introduced by Koren et. al 2000. Similarly the a-priori parameters for the river routing model can be derived from hydrographic properties. As part of the modelling system development phase, a GIS based semi-automatic system to derive all a-priori parameters was also developed to facilitate consistent parameterisation and faster model system implementation. The modelling system's output includes; snow accumulation and ablation, top and bottom horizon soil moisture, river discharge and overland runoff as well as soil temperature including soil frost and thaw depths.

HOPS soil moisture, snowpack and soil temperature evaluation runs have been conducted in Sodankylä, around FMI's Arctic Research Center. Hydrological (discharge and runoff) evaluation runs have been conducted in 9 watersheds within the geographic area of Finland. All evaluation runs were conducted without any calibration or deviations from the a-priori parameterisation schema. The performance of the modelling system was evaluated in all cases by running the model with historical meteorological input (daily precipitation sum, average daily temperature and sunshine hours sum) between the years 2005-2013, after an initial cold start period of 2002-2004. In all evaluation cases the HOPS system showed a high degree of correlation with the comparison variable observations. The largest discrepancies between the simulated and observed metric occur during snow melt periods as a result of errors in meteorological forcing data. To mitigate this, an approach to assimilate EO data based Snow Water Equivalent (SWE) data was developed. The assimilation of observed EO SWE data improves simulation results significantly during snow melt periods.

A stripped version (for simulating soil moisture and temperature) of the HOPS modelling system has been implemented to cover the entire geographic area of Finland with a resolution of 5x5 km. This model is being configured to run operationally to provide near real-time data production at FMI.

Potential applications: HOPS system is ready for a public audience. It has utility value on its own but it could also potentially provide the basic dynamics input for a Machine Learning based water budget model and soil mechanical model, both with resolution of 20x20 m. Grid of this size would make it possible to couple the relatively constant public data features and the weather forecast information to produce pointwise trafficability predictions. A lot of work remains in recasting soil models in the Machine Learning format and finding the potential verification sites.

4.5 Storm damage

A forest damage simulation was carried out in order to examine whether SAR data could be used to map storm damages to forests. Trees were felled in three different test sites near Parkano, and SAR images were acquired before and after the tree felling. Tree trunks were left lying on the ground until the end of the image acquisition. A comprehensive analysis about the state of the art in satellite radar methods applied to forestry is presented by Praks (2012) .

Results: The results indicated that X-band COSMO-SkyMed SAR data is not suitable for monitoring storm damage in Finnish conditions. This is due to the fact that the X-band radar reacts strongly to the branches and even the needles, and therefore the observed backscatter intensity (pixel brightness) did not depend on whether the trees were felled or standing. The same results were received when using SAR images in 5 m and 15 m resolution. Due to schedule limitations of Cosmo-SkyMed, a very high resolution (0.5 m) SAR image was acquired only after the tree felling, and therefore the analysis in that resolution was limited.

Potential applications: Experiments need to be made with Sentinel-1 5x5 m resolution images to see if the artificial or real damage sites can be detected. The advantage of Sentinel-1 compared to Cosmo-SkyMed with regards to forest damage detection is longer wavelength. Sentinel1 operates on C-band which is less sensitive than X-band to small particles such as needles and leaves . Instead, the C-band reacts more with the tree trunks, allowing for a better separation between standing and lying trees. In future investigation, SAR imaging direction with respect to the angle of the felled tree trunks should be also considered. In a real forest damage scenario this may prove difficult, due to the fact that the trees would be falling in different directions. Moreover, in real storms, fallen trees are sometimes uprooted and the lifted root system can be several meters in diameter, creating additional scattering surfaces. In simulated forest damage these do not exist.

The texture classification methods have to operate at a sub-optimal range. Thus one would need a lot of teaching samples from different snow and moisture conditions for the classification to be feasible. Also, the public data with large numbers of features has to be used to properly cluster the texture component. At the moment there is no direct applicability to forest damage. A new campaign is needed covering wide and diverse test area sets using the new Sentinel-1 as a source.

4.6 Recognition of mass-flow deposits for infrastructure construction

Mass-flow sediment morphologies are characterised as a series of ridges and mounds with varying elongation and sizes but always composed of poorly sorted sediments. Mass-flow sediment deposits most often occur as fields and are located in regional scale topographic lows. Mass-flow sediments (Fig. 4.15) are often poorly sorted diamictons but their content of fine grained fraction (clay and silt content, <0.006 mm) is usually less than 12% (Sutinen et al. 2009). Therefore, they are potential aggregates for infrastructure construction. Irregular distribution of boulders and stones in the top sediment sequence is typical of mass flow sediments. In addition, the surface of the formations is also often covered by an abundance of subrounded boulders of varying size. Therefore, mechanical crushing (i.e. crushing station) is required when mass-flow sediments are exploited as aggregate reserves, and oversized boulders cannot be utilised. Mass-flow fields such as the Kemijärvi field, in Finnish Lapland, has regional significance for aggregate production as there is an abundance of closely spaced mass-flow formations within a relatively short distance from the road network. The goal of this subproject was to semi-automatically map mass-flow deposits as potential new aggregate materials for infrastructure building with automatic pattern recognition methods using only airborne LiDAR data by the National Land Survey of Finland. In the first stage, all hummocky geomorphological patterns were recognised using pattern recognition of the LiDAR digital elevation model (DEM) and its derivatives. At the second stage, LiDAR point cloud analysis was conducted to recognise mass-flow deposits of the hummocky geomorphological units by analysing the surface stoniness. The study was conducted at the Kemijärvi study area where a large field of mass-flow deposits cuts across the study area.

Fig 4.15. The sediments of mass-flow deposits have low contents of fine grained matrix. The surface of mass-flow deposits are covered by boulders of various sizes and shapes. Photos: R. Sutinen, GTK.

Delineation of the geomorphological hill-like landscape features was done by applying a relatively new image processing and pattern recognition technique called Object-Based Image Analysis (OBIA). An OBIA algorithm applying Cognitive Network Language (CNL) in a commercial software called eCognition (Trimble Geospatial, Munich, Germany) was developed in the project to delineate the all hill-like geomorphological landscape units from the DEM. The commercial software was applied for this demo because the task is not easy to accomplish.

The algorithm first locates the hill tops and grows the tops 'downhill' in order to establish each hill as a separate geomorphological unit. The algorithm utilises raster processing of the DEM as a first step: tilt derivative (Miller and Singh 1994) (the arctangent of the ratio of a vertical to a com-

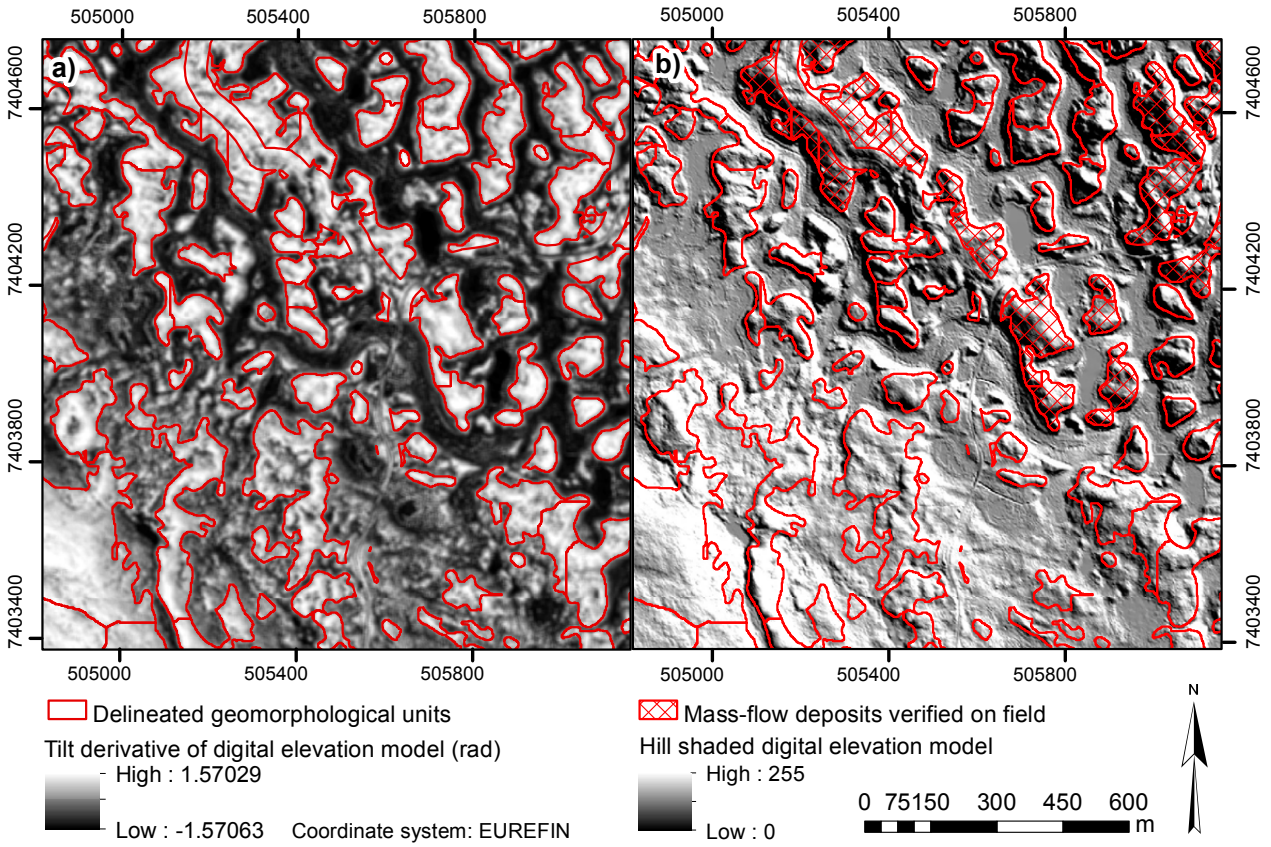

Fig 4.16. Subset of tilt derivative (a) and hill shaded digital elevation model (b) at the Kemijärvi study area. Mass-flow deposits are located in the central and north-eastern part of the map. The south-western corner is dominated by moraines. Figure by M. Middleton, GTK.

bined horizontal derivative) was calculated to emphasise the geomorphological features. Further processing was restricted to outside built-up areas and water bodies by buffering them by 25 m because aggregates under current infrastructure cannot be exploited for quarrying. The DEM was segmented by looping 'Multi-threshold segmentation' tool such that the pixel with the highest elevation was recognised and elevation contours were segmented in 0.5 m intervals at lower elevations until the lowest elevation in the study area was reached. Hilltop segments were then selected from these objects using a feature called 'Number of brighter pixels'. The 'flat area' class was created by 'Multi-threshold segmentation' of the tilt derivative (TDR). In theory, TDR value 0 appear at the edges of the 'hills' e.g. where slope of the DEM derivative turns from positive ('hills') to negative ('flat areas') TDR values. Finally, a two step growing procedure by the algorithm called 'Pixel-based object resizing' outside the created class 'flat area' was performed to establish the final boundaries for the hill-like geomorphological features. Numerous clean-up processes were performed as a preprocessing for these major steps.

In the second stage, four different methods were tested to produce the signal for the stoniness feature:

1. Local planar or bilinear fit by a nonlinear weight function to produce a local surface normal vectors, from which the likelihood of the presence of stones was computed by a modified Hough transform for the semi-spherical shape.
2. Digital Elevation Model (DEM) produced by public software. The reference raster length was set to minimum numerically possible (0.6-1.0 m depending on the local sample density). As in the first approach, a similar cumulation to normal vectors then followed.
3. Producing the alpha shape representation of the sample points. Alpha shapes have only one parameter, the radius r defining the maximum negative curvature allowed to the outer skirt of the alpha shape. The ground triangulation consists of the so called alpha-exposed vertices of

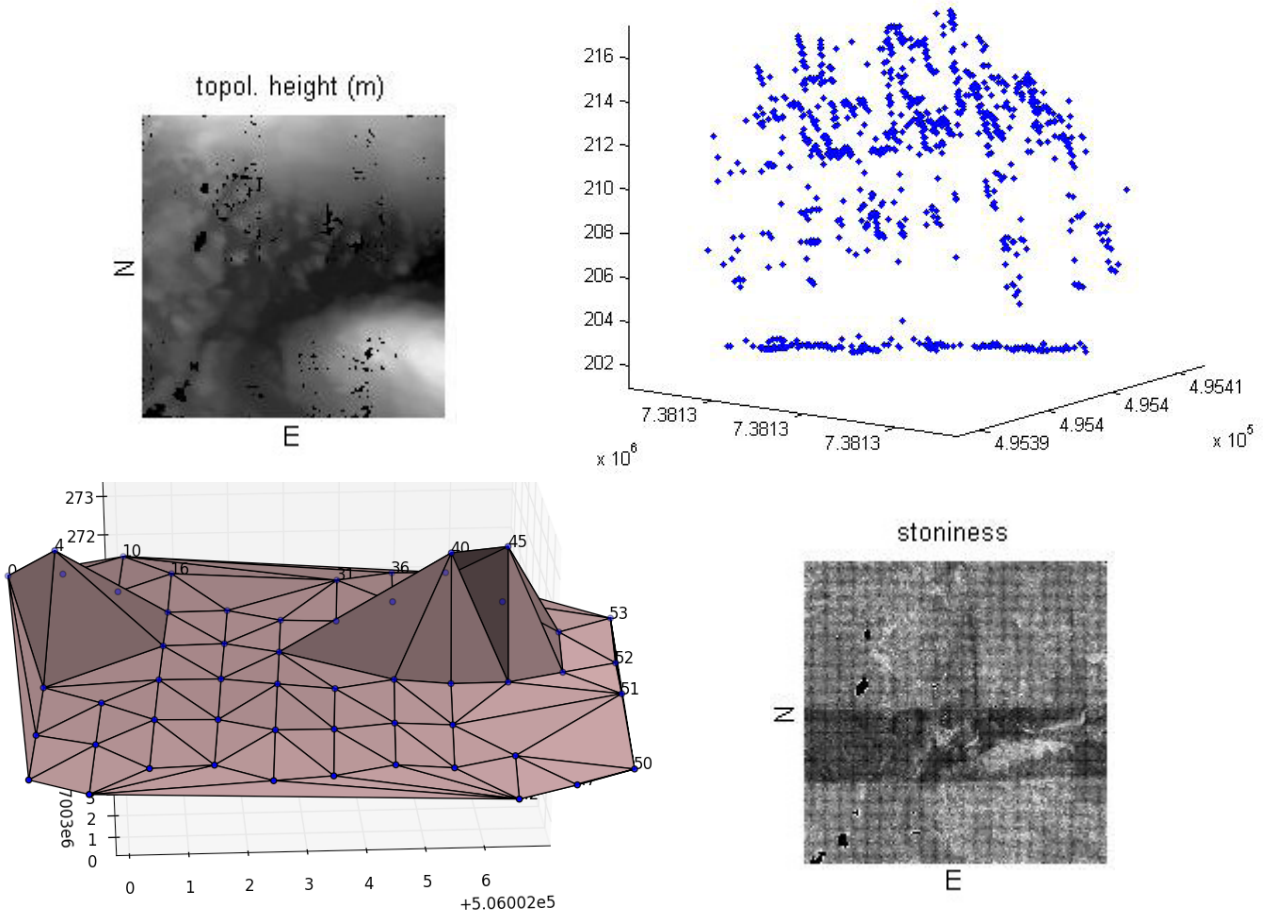

Fig. 4.17. Stoniness index at UTM map T5212D2: topological height (*upper l.*), LiDAR point cloud (*upper r.*), surface triangulation (*lower l.*) and resulting stoniness likelihood (*lower r.*). Figure by P. Nevalainen, UTU.

the triangulation. Experimental tests were made with radius , where the lower limit was clearly too small. The radius parameter of the alpha shape defines the smallest size of stones where the semi-spherical Hough transform could fit a stone. The rest of the analysis was like above.

4. Subjecting 2D projected points to Delaunay triangulation and eliminating the sharp vertices by a spatial angle filter. The third image in Fig 4.17 (*lower left*) depicts two stones, which are later detected by a special voting procedure.

Fig. 4.17 depicts aspects of the stoniness computation. *Upper left*: one 1.5 x1.5 km quadrant of the UTM map page T5212D2, topological height derived from the point cloud surface triangulation. White corresponds to higher elevation. *Upper right*: LiDAR point cloud, the tree masses above and the ground surface are clearly visible. No additional information (return intensity, number of return signal etc.) available in the raw LAS file format (.las) was used. This is because there are a lot of data collections available without this information. *Lower left*: method 4 described earlier in the text was used to produce terrain triangulation with the minimum possible loss and noise. *Lower right*: The likelihood of the presence of stones computed based on method 4. White corresponds to higher stoniness. The grid pattern is an artifice of space partitioning. The vertical stripe is artifice of doubled sample density on a zone, which has been scanned twice. None of the methods presented can be made sample density invariant at the low end of the density (), they are sample density invariant at the higher densities, though.

Results: Fig. 4.16 (*left*) illustrates the potential mass-flow segments that were a result from the object based image analysis. The algorithm delineates the outer boundaries of the geomorphological units well. However, further development work is needed to resolve the boundaries between each sub-polygon within each larger geomorphological unit. Two full days of field work was done to select training sites of the delineated hill-like polygons for mass-flow deposits (56 objects) which were recognised by having surface boulder cover and non-mass-flow deposits (52 objects). In figure 4.16 (*right*) mass-flow morphologies are situated in the NE corner of the map. In future work, the surface stoniness of these geomorphological polygons has to be examined in order to classify the hummocky geomorphologies into mass-flow deposits and non-mass-flow deposits.

Method 1 gives the highest quality for the surface stoniness but requires the most computing, while method 4 seems most plausible for use in stone detection. It is rather fast with time complexity of applied to patches of appr. n= 5000 points per time. It detects stones more reliably than public GIS software (method 2) in the size range of stone radius 0.8-1 m using the sample density available (). The quality improvement has not been published yet, though.

Potential applications: This method has high relevance in mapping these previously unrecognised aggregate deposits i.e. mass-flow formations on all formerly glaciated terrains. The work is the first attempt in Finland to map geomorphological units semi-automatically using Machine Learning methods on LiDAR data. The utility value of this approach would be high in many fields of geosciences, e.g. aggregate mapping e.g. for the arctic infrastructure development, terrain trafficability for forestry and military purposes, and geomorphological research. The implementation could be conducted in two phases:

- delineation of hill-like features from LiDAR derived DEM which could be conducted as a batch run from tiles of the nationwide data.
- estimation of surface stoniness which could also be a one-time run to all available LiDAR data in Finland after necessary scientific validation processes are done.

The method of recognising stoniness from LiDAR point cloud is expected to work especially well on terrain with sparse tree cover such as in northern Finland. Forests with dense canopy may obscure the ground signal beyond the practical sample density limit. Another hindering issue is generally denser field vegetation in the southern boreal forests. LiDAR data with higher point density are becoming economically feasible, and the Machine Learning methods could be developed for the situation where the ground signal is available only from fragmentary parts of the forest. This will be a subject of further research in study areas where field data on stone coverage and size distribution will be available. In this approach, where the ground truth comparison is used directly, a new field campaign is inevitable.

The application uses only the ALS data by NLS which will be available nationwide by the end of the year 2019. In order to recognise the surface stoniness it is recommended that further ALS data acquisition by NLS would be done using higher point density, especially in the southern boreal forests. Besides the application of surface stoniness presented in this study, stone density and size estimates would also be important inputs into the terrain trafficability models (Sec. 5.1).

5 TECHNOLOGICAL POTENTIAL

The study participants have gained a solid understanding of potential future applications. Ch. 4. has short presentations about some of the most promising applications next to each test case. This Chapter concerns implementation issues, computational formulation of the problem and levels of integration with the existing GIS software. Sec. 5.1 is about implementation, Sec. 5.2 lists the computational tasks, Sec. 5.3 concerns the workflow issues, and 5.4 deals with the cloud aspect. An introduction to ULJATH implementation aspect can be found in Nevalainen (2014a).

5.1 Implementation

The Machine Learning problem matrix structures (Eqs. 3.1 and 3.2) are presented, since they drastically differ from the usual approach of integrating algorithms into existing GIS software. Since the problems can be huge, matrix **M** has to be generated column by column from GIS platforms and the result saved outside the GIS before the Machine Learning codes can start the analysis. This is reasonable, since a large set of ready-made software accepts matrix **M** format, while only a few integrate readily into GIS, and those which do generally have memory problems. Also, many computational tasks are cumulative concerning e.g. the whole nation, therefore managing suitable batch processing files is more important than the personal GUI aspects and easy visualisation which GIS provides.

GIS provides some tools for locational data management. The following table lists the initial data handling tasks and their potential implementation site. The following shorthand names are used:

- **PIG**: (Platform Independent GUI), an applet etc., accessible by wide public
- **PC**: usual personal computing, which is completely at command of the user:
- **GIS**: the corporate installation of GIS with the usual interfacing to wide-scale databases or possible organisation-wide cloud.
- **cloud**: cloud envvironment

Table 5.1. Data flow, tasks and possible implementation platforms of tasks. Data flow direction expressed with arrows.

	task	PIG	PC	GIS	cloud
1	primary data upload			X →	X
2	area selection		X ←	← X	
3	small scale sampling (quality assertion, design)		X ←	X ←	← X
4	individual batch task design (algorithms)		X →		→ X
5	batch flow design, test runs		X →		→ X
6	window sampling, map tree control, building **M**				X
7	prediction initialisation (train-validate-test)				X
8	results: predictions and forecasts		X ←	X←	← X
9	end user products	X ←			← X

Table 5.1 is commented in the following. The number indicates to which specific table row the comments apply. Note that not all rows are commented upon.

1: Primary data upload occurs seldom for most producers, e.g. Metla NS-FMI data (forest inventory) is updated biannually. GTK data has a similarly slow cycle, yet the geological data is likely to become expanded and more varied over the years. FMI data (meteorological nationwide grid and the 3-7 days weather forecast) should be updated once or twice a day nationwide. In the long run it is reasonable to have this data in the cloud within the donor organisation and then transfer it between two cloud systems. The difference between the data representations of two cloud systems, the donor organisation (GIS) and the proposed central system (cloud) is that the latter enables direct Machine Learning processing. The data is in the data matrix format (matrix **M** of Ch. 3). Mapping between two formats is rather simple, but numerically inefficient if done each time a different computational task is entered.

2: Area selection and the corresponding data cuts are typical GIS functionality which usually only require a minimal set of shape information (rectangle, circle, polygon, crossline), and can be manually added to the batch processes. We define a batch process as a computational task which requires a human decision to commit to. GIS is useful here, but only for producing the shape information. Actual data cuts and merges need not happen within GIS.

3: Current GIS systems excel at quick sampling and cartographic visualisation of experimental and intermediary results, in addition to the production of documentary material. The actual computation should be performed using either a PC or the cloud, using the standard data matrix format. This removes the need to repeatedly add experimental code to the GIS software.

4–5: The cloud batch files are easy to develop, experiment with and verify on PC, when one limits the analysis to moderate test problems. Relying upon standardised batch file conventions and omitting GIS in this phase simplifies the task.

6: By moving a limited set of core GIS functionality to the cloud, one gains orthogonality between two environments (GIS and cloud). The suggested functionality is in the colored box.

7: A typical ML See Ch. 3

8: Analysis of results can be performed using numerical software or GIS. This phase is mostly for methodology development and not part of a final product.

9: GIS functionality is already integrated and hidden in many modern locational data applications, see e.g. GIS Cloud. One can expect that the applications in the near term future can be based on third party application platforms providing high level GIS+cloud integration and support for distributed applications and architecture.

The point of the Table 5.1 is that a large part of processing can and should be made outside the GIS systems. It is probable that all main components of the intended system (PIG,GIS,Cloud and Machine Learning algorithms) will continually evolve in the near future, and the proposed policy of minimising GIS dominance may help in three ways:

1. the adoption of emerging technologies will be easier, since proposed components are orthogonal
2. research problems can be widely disseminated, and a large segment of professionals can participate in developing new applications
3. expansion of the methodologies can happen independently amongst data providers, GIS tools, Machine Learning methods and developers of commercial applications.

5.2 Computational tasks

This project is limited to the static data with updates occurring biannually at the most. It is worth noting that there is much more static data when compared to dynamical input (weather parameters). The following calculation demonstrates this:

Table 5.2. Rough estimate of data amount of a potential national service.

amount of grid points	non-water-area / (20 x 20 m^2) = 800 x 10^6 = n
amount of static data (floating point numbers FPN). d_s = 90 is the number of features	nxd_s = 0.7 x 10^9
m_0 is amount of weather data observation points	non-water-area / (150 x 150 km^2) = 20 = m_0
dynamical input data stream per week in FPN. d_d = 30 is number of dynamical features, is the number of weather report inputs per week. The constant 2 takes into account that both weather forecast and the weather report are fed in at the same time.	m_0 x d_d x 3x 7x 2 = 25 x 10^3 = m

This disparity between the size of static and dynamical input means that rather heavy costs can be accepted with analysis of the static part. Dynamical time ticks can occur 1-3 times per day as long as the computation can be finished within one tick. The consequences to Machine Learning algorithms are not considered in this report. The expected dynamics of the water budget model are relatively small - not many iterations are needed per time tick and location before a new equilibrium is found.

Computational tasks are either local (initial area and data selection done in an interactive way) or in the cloud (batch processing). The computational tasks are listed here in order, for later detailed analysis with reference to the batch processing layer. Some of the tasks can be done in GIS but national applications need such a scale of automation that GIS might be used for preliminary tests and runs and to quality control and not for actual batch processing.

Pre-processing:

- Spatial cuts and joins along the GIS tradition taking into account the varying sizes of the feature maps. Raster data is modular to UTM naming system, and cuts and joins of these is a typical GIS task. But as a numerical preprocessing step, this task has to be implemented in batch processing level as well.
- Eliminating or imputing invalid values. Imputing is valid in cases, where there is an established correlation between the missing feature and other available features. The effect of imputation has to be checked by the final performance criteria in each case.
- Changing the grid sizes (both towards higher and lower density). This is done by localised radial interpolation and in opportunate cases, by bilinear interpolation.
- Turning the raster information to data matrix format (see Eq. 3.2)
- Regularisation to $N(\mathbf{0},\mathbf{1})$ distribution or normalisation (to unit norm)
- Interpolating the features to the same grid.

Generating the derived features:

- Image processing, Machine Learning (ML). Methods implemented in Matlab and python.
- Derived features related to geographic height. Methods implemented or available in Matlab and GIS software.

Generating the actual predictor:

- Generating altered ('dead-zoned') training sets: matlab
- Ridge regression, k nearest neighbours (k-NN), multi-layer perceptrons (MLP): matlab

Thinning and feature selection: Feature selection was tested in this project, but neither it or thinning has been implemented in this project. These tasks are important, and they are being discussed more in Sec. 5.3.

Visualisation: This project has only static images as the end result. Combining subareas to an entire map and enabling typical browsing operations is a typical GIS task, but it must be governed by batch processing tools too.

5.3 Review of GIS workflow

Geographic Information System (GIS) is often understood as a system to create, edit and store geographic information, i.e. information which has a location as a central integrating concept. Most commonly geographic information processing requires direct user interaction via GUI. At some cases, user can record a sequence of actions to a batch file. This batch run can be performed to a larger geographic area. Usually the batch file requires expert modification and integration with third party libraries.

There are several commercial software for semi-automated and automated GIS-analyses, such as ArcGIS, Erdas Imagine, MapInfo and Smallworld. During the recent years open-source code software have become increasingly popular. Few notable examples are QGIS, SAGA GIS and Grass GIS. QGIS is particularly interesting since it incorporates other open-source software code and tools into its GUI.

Many types of software also include a built-in scripting interface where it is possible to customise workflows and even build new tools within a program. Python scripting is probably the most commonly used language in user scripting. Most of the GIS-programs themselves are built using C or C++.

Automated workflows can be created with user scripting. Modules or tools can be run for example through python code. This way it is possible to link different tools into one script and it is not necessary to do all the analyses separately. This type of approach was used for example in the flood mapping demonstration and in the drainage network detection. The advantage of this approach is that it is extremely flexible and it enables fast and repeated processing of long processes. The only limitations are tool availability and the users coding ability.

Commercial softwares have visual automatisation tools, such as the ModelBuilder found in ArcGIS software or the Model Maker found in Erdas Imagine. These models can be saved, modified and re-run. A similar example from open source software is the SEXTANTE model builder. We mainly used user based scripting to automate the workflows in SAGA GIS and Lastools. ArcGIS ModelBuilder was utilised in few examples to automate a few individual tasks, however it is not used in any particular demo.

Online-GIS applications have also become increasingly popular in recent years, but so far the inability to complete complex analyses which are possible in desktop-GIS excludes online-GIS from most analysis tasks. However, this field is developing rather quickly and it is likely that Machine Learning or other independent numerical analysis is processed at cloud only. In any case, we believe that desktop GIS will be an important part for a long time, at least in more specialised analyses. After new algorithm development and testing, and if the resulting new feature is deemed useful, a batch process can be launched in the cloud server to produce said feature at regional or national level.

5.4 Cloud environments

The concrete outcome products of this project may be included in the FMI lead national satellite data center archiving, processing and distribution system installation, which is currently under development in Sodankylä in Spark and EnviBase+ projects. The thematic information required in future national resource utilisation planning will be based on numerous new data sets which will be combined for new applications under development. This requires a combination of new computational and numerical data analysis and deep understanding of the natural processes being analysed and estimated.

5.4.1 GUI

One GUI experiment was performed using the Amazon EC2 cloud service and Anaconda ipython by Continuum Analytics. These products were selected due to the ipython notebook functionality and wide sortiment of GIS -related Anaconda libraries. We developed a simple protocol to transfer

map layers, i.e. locationally bounded feature sets, and view them. GIS map merge and view functionality and rectangular cut was implemented. As a test application there was an expert rule of thumb for finding potential unconventional mass-flow deposits:

pine volume per hectare AND sandy till soil type
→ potential mass-flow deposits

The expert rule is intended for Northern Finland, but here it is applied to the Parkano area for demonstration purposes, see Fig. 5.1. This early test was meant to demonstrate the technology and it predates a more detailed study done in Sec. 4.6.

The response times seemed acceptable, although the test was performed with a limited set of map pages. Most of the components of the software have recently become available as open source. The complete application would need three more components: the road network, topological height and the surface stone coverage estimate. The latter would be one more feature to the expert rule, and two others would enable grading potential sites by their economical utility (transport distance) and environmental aesthetics (extraction site visibility). Even this kind of deliverable would have been of high interest, it would have been difficult to verify its accuracy and impact scientifically. Thus only a modest demonstration was done.

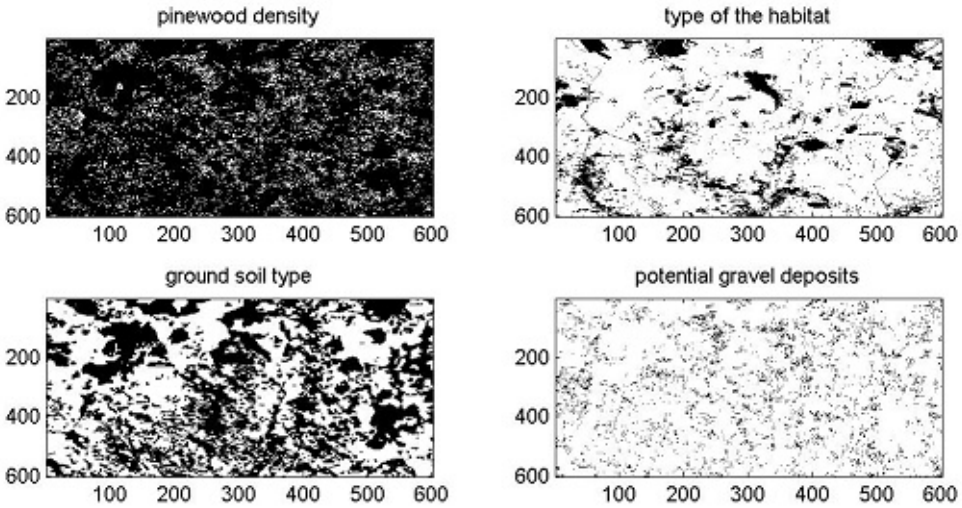

Fig. 5.1. Three known attributes combined by an expert rule to likelihood of potential gravel deposits. The highest likelihoods for each input variable are depicted with black colour. The low right figure depicts the predicted gravel deposits, the final output, also with the black colour. Figure by P. Nevalainen.

5.4.2 Cloud providers

Only the most prominent providers are mentioned, and they do not all compete on the same product and service sector.

1. National satellite data center: (FMI, Sodankylä) is a strongest contender when considering the implementation for the natural resource open data server.
2. Techila Oy: Techila's cloud integration solutions on the field medical imaging are rather close to the needs of this project. Techila could suit in cases where data is partly public, partly private.
3. GIS Cloud: This Slovenian company has a high-quality set of tools for user interfacing to large and distributed systems using variety of technology. The customer-based field data gathering is also well implemented.
4. Continuum Analytics: Their cloud solution allows easy local development, testing and quick deployment. They have good integration of scientific python tools including GIS libraries to their Amazon EC2 cloud solutions. Their approach is suitable for small projects which exceed the capacity of existing GIS environments. Also, they have merits what comes to incomplete data processing, see e.g. Darpa Xdata project. They have several contracts in place for various geological and explorations firms. Especially Anaconda Cluster seems to be an environment of choice when research has to be applied rapidly to real scale problems.
5. Whitebox Geospatial Analysis Tool: extensible, can integrate various research level algorithms. Their tools are good for education and tutorials.

6. Others, e.g. Tieto Oy, DataCenter Finland Oy. The scope of this project does not allow complete analysis of the cloud provider sector.

As a general note, most of the Machine Learning project environment providers and cloud providers seem to acknowledge the importance of smooth workflow from expert crafted algorithms to system integration emphasised in (Rehr 2012). GIS field has similar approach only for platforms for mobile applications, see GIS Cloud (the list above), and in some extend with SAGA GIS tool.

6 DISCUSSION

Sec. 6.1 presents deliverables of the project and Sec. 6.2. the summary. Sec. 6.3 includes a summary of possible future research and there is a short survey of commercial applications in Sec. 6.4, since the various activities of this project will continue among the participants after this project.

6.1 Deliverables

The project has delivered the following concrete results:

1. Hydrology prediction system (HOPS, see Sec. 4.4) is already functional and can be implemented e.g. at the Sodankylä satellite data center. HOPS can feed the essential dynamical input to the water budget model and to a future trafficability prediction system. HOPS will be published separately by MFI.
2. Case studies which are close to real application status. The only requirements are a customer-provider infrastructure and some software integration. These cases are listed here in order of decreasing maturity:
   a. drainage network detection in low-relief forested areas (4.3)
   b. an application for flood detection in forested areas (4.2)
   c. processing algorithms to delineate geomorphological units and to estimate surface stoniness from ALS data in order to map unconventional gravel deposits (4.6,5.2.4).
3. Prediction models from the ML test cases, see Secs. 4.1.1,4.1.2,4.1.4,4.1.5 and 4.1.6. The models will be made publicly available for the research community. Included will be: data matrix **M** (see Eq. (3.1)), the list of features (columns of **M**) and a short background summary for each test case.
4. Specification of the hierarchical data structures and batch processing arrangements/policy needed to produce these models in a generic, nationwide setting. The presented policy is aimed at local development and testing of computational methods, which are then installed in a cloud environment in a straight-forward manner.
5. Several public presentations, two scientific publications (Pohjankukka & Nevalainen 2014, Pohjankukka 2014) and two M.Sc. theses (Pohjankukka 2013, Nevalainen 2014b).

The results mentioned in items 3,4 and 5 will be made publicly available in 2015. The actual project archive site is not yet known. Results from the cases in Secs. 4.2, 4.3 and 4.6 will be sent to scientific journals during spring 2015, once the analysis and writing are complete.

6.2 Summary

In each test case we noticed that acquiring the field reference (a.k.a baseline) required resources and careful planning. Methodology validations relying on the public data itself were omitted. There is further discussion in Sec. 6.3 about methodology validations without field campaigns.

In the trafficability analysis we focused on summer and autumn trafficability. That way, the validation of crucial elements from a future trafficability system was possible. The proper scientific and practical approach for coupling the thaw model and spring time trafficability conditions requires a separate effort.

The shift from the Cosmo-SkyMed satellite imaging resource to the Sentinel1 changes the bandwidth from X- to C-band, and therefore makes the nationwide time series analysis of satellite images possible. It is probable that the time series approach improves analysis based on satellite SAR technology. The time series approach with Cosmo-SkyMed was economically unfeasible even using limited target areas.

The results from each case were presented in Ch. 4. The following is a list of general results and insights gained from the project:

- Relating public data features with a specific physical phenomenon nationwide while achieving high prediction performance requires extensive field references for validation of the models products. Optimising such campaigns, i.e. selecting the field sites, requires an existing nationwide data service and preliminary analyses.
- Positive results were achieved with the drainage system evaluation, route planning and flood detection. Almost all applications require good ALS data. The existing NLS ALS service provides an excellent base, however we stress the need for frequently updated ALS data, preferably with higher pulse density than currently. This would benefit many applications via enhanced target signals, e.g. ditches and boulders, and may open completely new possibilities for further development of applications.
- The forest harvester route choice test based on Pieksämäki soil damage data (see Sec. 4.1.6) provided good results. It is likely that other applications based on a similar cumulative indicator will perform well with similar types of data.
- Some properties essential to trafficability prediction e.g. hydraulic conductivity, soil bearing class and penetration resistance (see Ch. 4.1) show strong spatial dependence and weak generalisability according the used ML methods. This means that future applications need an online learning approach unless new, useful data sources become available or a breakthrough happens in derived features generation and environment classification. One possible advance in trafficability prediction could be achieved by the utilisation of vertical distance in the drainage network. A comprehensive drainage network can be generally well retrieved from ALS data as shown in sect. 4.3.
- There will be full ALS coverage of Finland in 2019. Drainage system analysis is an application utilising the LiDAR data which has direct commercial potential. The analysis can also proceed on a case-by-case basis and is thus independent of the national data resource.
- Geomorphological mass-flow features and geomorphometrics are an interesting new aspect if subjected to the ML approach, and they may provide additional information for environment classification. Since geomorphology is a logical aspect of the natural environment, there could be an advantage over conventional multi-feature and large-window texture analysis.
- Separate experiments about sample thinning (see Ch. 3.3) among three test sites (Parkano, Pomokaira, Pieksämäki) show that significant sample size reduction could be possible before the usual application of Machine Learning methods. Thinning is an orthogonal aspect to feature selection, which is a common technique in ML, but which did not promise the possibility of large computational savings. This is another favorable point for the nationwide data service.
- Economical feasibility of other deliverables is hindered without a national public data resource. Commercialisation requires some protection from ongoing organisational and technological changes among data providers. The data resource service itself is an unlikely subject for private entrepreneurship, since the investment return time is potentially quite lengthy.

Collaboration between the project parties has provided a useful impetus towards future projects.

6.3 Future research

In the future some research has to be directed towards methodology validation by means of widely available data (public data itself) without expensive field campaigns. This kind of research can be performed on a low budget and can happen between major projects. Subjects of methodology validations include the generalisation capability, potential of thinning the sample sets, feature

selection, more powerful texture methods, and a wider assortment of Machine Learning algorithms.

Another research arm is generating a multitude of models related to general trafficability. These models include:

- Hydrologic prediction system (HOPS, Sec. 4.4), which couples weather data, local large scale geographics, and evapotranspiration models. It also initialises a possible thaw model.
- Water budget model, which is a high-granularity version of HOPS, taking into account small scale topology, local soil properties and has an interface with the soil mechanics model.
- Soil mechanics model, which estimates the current load bearing capacity at a given point. This model needs to also access the available and predicted soil properties.
- Thaw model, which combines two previous models with meteorological data, and is applicable only during the winter season.

Several small scale research projects can prepare for future unification of these tools in order to achieve generic trafficability forecasts. The grand unification of all tools will require careful preliminary studies, planning and wide co-operation.

Any future application would need at least some of the following data sources provided nationwide in a unified manner within ETRS-TM35FIN coordinate system. Data sources currently unavailable but possibly available in the future are marked with an asterisk (*):

1. Existing features (see Ch. 2) based on public data directly used in this project.
2. Regularly updated Sentinel-1 SAR images for 2-3 different seasons. A similar arrangement for visible wavelength, snow water content and snow coverage.
3. Relating the meteorological forecast of several dynamical variables, e.g. precipitation, evapotranspiration, temperature and snow water content to general trafficability models.
4. Dynamical soil model, which includes mechanical strength, water budget and thaw models. (*)
5. Thaw model, including the prediction of the end of the thaw period. See (Sirén et al. 2013) about the importance of the spring period in forestry transportations.
6. Geomorphological features (*)
7. Additional measurements, which are needed for specific applications:
   - transmission power expenditure and wheel sinkage of harvesters and forwarders(*) (Metla).
   - ground radar results with varying weather history and from nationwide measurement sites (*) (GTK).
   - forest ground ALS data with at least 2-4 points per square meter.
8. Texture analysis library, which can detect textural patterns over large window frame (100x100 pixels) and has support for a crucial set of invariance properties (rotation, satellite image inclination, translation, gray-scale) (*)
9. National natural resource public data storage serving research and commercial applications (*)
10. Modern multisource features. An example is using plant and vegetation recognition based on hyperspectral remote sensing (Middleton 2014) to feed a further texture segmentation phase.

The conclusion of Ch. 5. is that technology already exists for providing many of these facets in an organised manner and on a national level in the future. Based on that, the following nationwide applications are possible:

- one week trafficability forecast for forest machinery and logistics
- pinpointing unconventional construction material deposits while taking into account logistics and nature conservation
- automated assessment of forest drainage network
- increased accuracy of flood prediction

If the ten data sources listed in this subsection could be fine-tuned using Finnish sub-arctic areas (northern part of Lapland), then the geographic application range would span over global arctic areas, thanks to the rather homogeneous character of the polar zone.

6.4 Future commercial applications

Four cases are presented. First there is a short discussion of the prerequisites for future application. At least some of the 10 facets listed in Sec. 3.3 have to be provided nationwide in a unified manner before further commercialisation.

The conclusion of Ch. 5 is that ten aspects mentioned in Ch. 6.3 are technologically feasible and can, at some point in the future, be provided nationwide in an organised manner. Using this assumption, the following applications are possible on a national level:

- One week trafficability forecast for forest machinery and logistics
- Pinpointing unconventional construction material deposits while taking into account logistics and nature conservation
- Automated assessment of forest drainage network
- Increased accuracy of flood prediction

If the ten data sources listed in this subsection could be fine-tuned using Finnish sub-arctic areas (northern part of Lapland), then the geographic application range would span over global arctic areas, thanks to the rather homogeneous character of the polar zone.

6.4.1 Terrain trafficability forecast for forest harvesting machinery

Trafficability modelling of the unpaved terrain requires weather forecast input and an integrated water budget model. The latter is complex and requires adaptation to be trafficability specific. Presently, the problem of coupling the water budget to public data in a way that the system can be subjected to Machine Learning methods remains open.

Fig. 6.1 presents a simplified systemic view of the possible system. The emphasised boxes are already implemented. Another open research problem is to design a field campaign to properly verify the theoretical model and to arrange the Machine Learning based fine-tuning of the model and verify it. The process will be tedious and elaborate due to the following reasons:

- Topology couples dynamics of distinct locations, thus field observations have to cover large areas
- The local water saturation level is a dynamical entity with possible causal effect over an interval of 6 months - thus long observation series are needed
- Local water saturation level responds to daily precipitation quickly, thus observations need to be dense
- A valid teaching sample is in the domain of time period x area

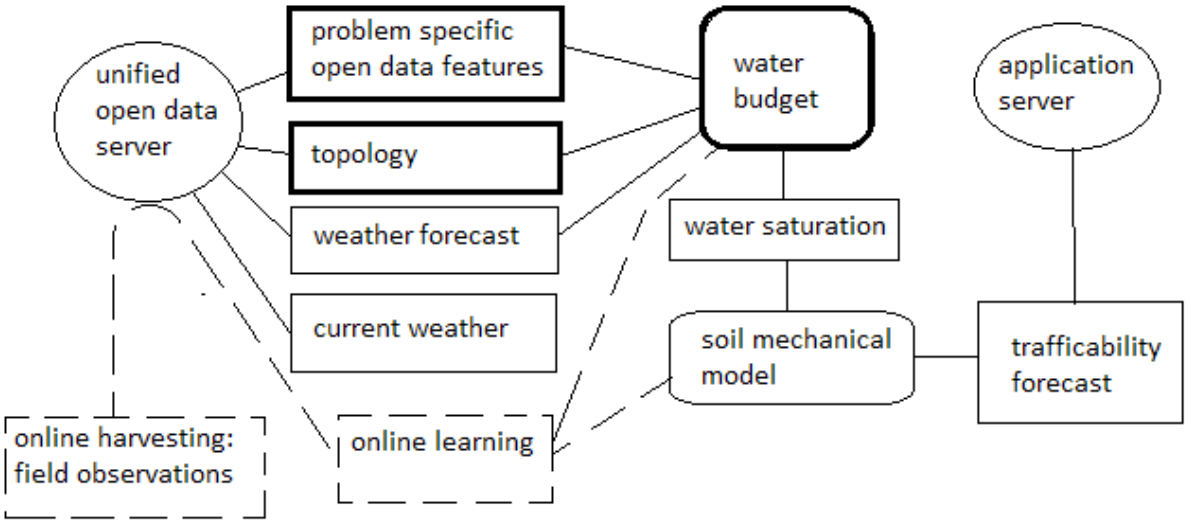

Fig. 6.1. Water budget model and trafficability forecast. HOPS system (see Ch. 4.4) would transform the weather forecast to localised water budget model input. Figure by P. Nevalainen, UTU.

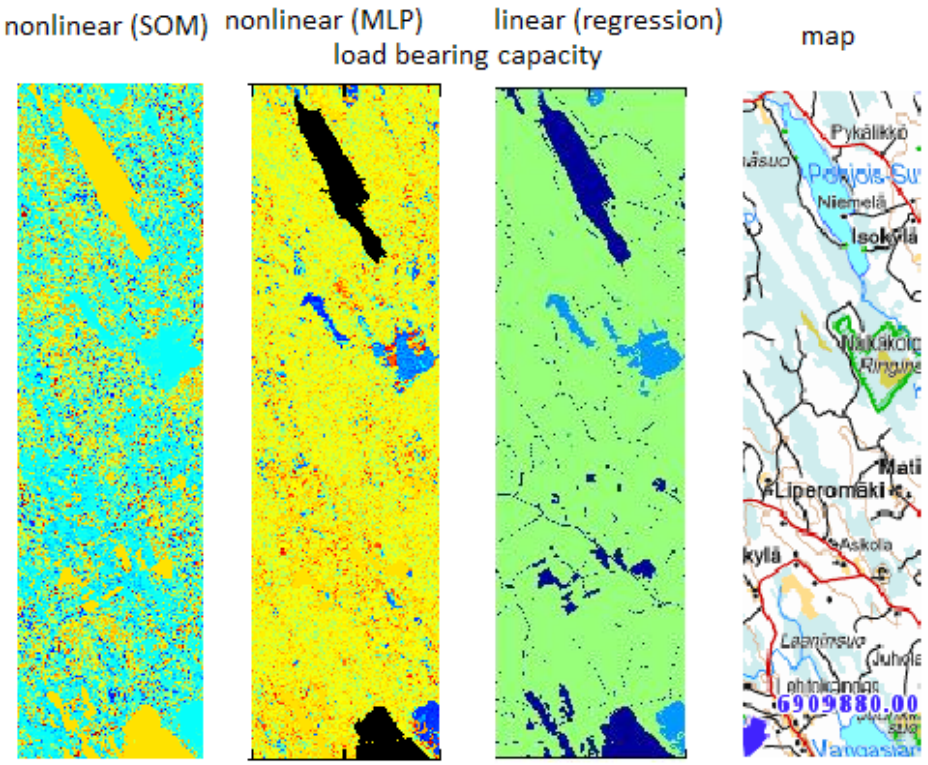

Fig. 6.2. A preliminary rut damage generalisation. Dark blue and pale blue correspond to the highest rut damage potential. Figure by J. Pohjankukka & P. Nevalainen, UTU.

Given the water budget model, the trafficability forecast would be technologically feasible in Finland and Arctic areas as a weekly forecast. The water budget can be described as a dissipative system with an integrative element. The basic idea of the model is that cavities in the granular structure serve as a reservoir for water. The water saturation state usually follows the precipitation, which is the driving input signal. The flood situations are exceptional and could be modeled in a different model with a greater weight on the direct interplay between surface flow and topology. These kind of models already exist in many mountainous areas.

The forecast map could look like the soil damage example in Fig. 6.2. Soil damage is related to general trafficability. Three different methods were used to classify the damage in new areas. Damage classes are: 1: no damage (green in the regression image), 2: moderate damage (blue in the regression image), 3: serious damage (deep blue in the regression image). There is no means to verify the prediction, since the actual field measurements cover only 1.5% of the whole area. There is a possibility to estimate the prediction performance though, in this case the predictions between 1 vs. 2 and 3 are ca. 58% correct, which may be adequate for route planning. The presented results are for visualisation purposes only, better results require addressing other aspects of trafficability (e.g. hydraulic conductivity, penetration resistance and load bearing capacity).

Rough estimates of the computational load for nationwide forest trafficability service are given in Ch. 5.

Commercialisation model: The concept could be a national level service for forest harvesting entrepreneurs on a weekly level and as needed for parties transporting heavy equipment in cross-country conditions (powerline construction and maintenance etc.). A large initial investment is required for learning the system and computing the model so that the dynamical input signal (weather data) can be entered. After that, the computational load is rather small concerning the update of the forecast on areas actively queried by customers twice a week. Twice a week is just a current estimate and serves as a basis for costs analysis. The final frequency can be optimised based on the forecasting accuracy and economic factors involved.

Properly launching this service may need an initial consortium between several interested parties including research organisations, universities, forest industry, Metsähallitus and forest machine manufacturers.

6.4.2 Forecasting forest road trafficability

This application is similar to the previous one, except some of the data features are different:

- The actual state of the forest (mass of branches, the average age of trees, the root quality) does not play a role.
- The drainage system arrangements near the road and road culverts dictate the water content of the road basement.
- The road structure is specific to the road category and transportation condition is strongly dependent on the inclinations occurring along the route

The data acquisition method is also different. A small all-terrain vehicle-carried ground penetrating radar (GPR) measures the soil thickness and also reveals the road cross-profile properties.

The total result of these differences is that it may be hard to implement the road trafficability forecast as a mere side product of the forestry off-road trafficability service. Road trafficability requires independent data sets and training, although the large scale water budget model can be common to both services.

Commercialisation model: Very similar to the previous one, except the computational load is likely much smaller. The soil mechanical model (see Fig. 6.1) is evaluated only at the grid points adjacent to the road. The customer base would be much more varied, and this application could be easier to launch.

6.4.3 Using harvester CAN-bus data for trafficability mapping

Presently the trafficability of a harvesting site is assessed by estimation of the foreman or the machine operator. Estimation can be difficult, but the present labour intensive pointwise means of measuring are too expensive. The target must therefore be set to comprehensive, continuous and low-cost assessment of trafficability by measuring. These requirements can be fulfilled using the harvester, which precedes the heavier forwarder, to create a mobility map of the extraction road network.

Harvester motion resistance can be measured using the data in harvester CAN-bus (Controller Area Network). At a steady speed on level ground engine power via transmission is expended on overcoming motion resistance, mainly dependent on wheel sinkage, which in turn is dependent on soil strength vs. loading. To determine expended power, the pressure and flow rate of transmission hydraulics and transmission rotational speed are obtained from the CAN-bus, and ground speed and terrain inclination are determined by auxiliary measurements. Further, data on the maximum available tractive force is needed to calculate net tractive force.

In practice the system would consist of detecting soil bearing capacity during harvester work and generalising the information for the surrounding cutting area based on independent ancillary information as demonstrated in this project. The application is hindered by the inadequate amount of field observations for research and development. The required information could be acquired by equipping a few harvesters to collect the data during routine harvesting operations.

Based on this concept, the measuring system would in the future be directly incorporated into the harvester by the manufacturer and the measured data registered by the harvester would then be available via a cloud server. This would enable the use of the information from the surrounding areas in modelling, not only for harvesting but for more detailed planning of other forest operations such as regeneration and scheduling the silvicultural operations. The described system could reach wider spatial applicability through an online learning arrangement, where the national trafficability model would be updated by each new observation. The same or similar terrain spot would yield useful information if weather conditions are different.

Commercialisation model: This application requires intense co-operation with key forest machinery producers, software developers and research organisations. A crucial part of the system is embedded software technology. The application has international potential both in civil and military usage, but in either case, the same long term coupling between machinery producers and cloud technology providers exists.

6.4.4 Thaw model

Thaw model alters the water budget model and soil mechanics model in two ways:

- Thaw state and amount of snow dictates the water budget initial state in the spring season. This has an effect on foresting operations during next few months.
- Thaw state and amount of snow dominate the trafficability during the winter seasons

Remote satellite observations about snow coverage and water content in snow have to be integrated with the basic feature set used in this project. This will feed the thaw model, which would produce forecasts and reports about various trafficability aspects.

Commercialisation model: The core set of customers would be the forest industry parties and agriculture, but there could be hitherto unseen revenue generating value for various freetime activities too. Many of the applications may spur general interest and provide goodwill for the participants.

ACKNOWLEDGEMENTS

"New computational methods for efficient utilisation of public data" project was funded by the Finnish Funding Agency for Innovation (Tekes). Special thanks go to Heli Laaksonen (NLS), who provided crucial answers to our queries. Field experience by Juha Majaniemi, Ilkka Aro and Markku Virtanen (GTK) in the RSAD surveys are highly appreciated. Occasional encounters with many experts of their own field has not been forgotten. Thus we mention names of Timo Ryyppö (FMI), Olli Sirkiä (NLS), Pekka Naula (UTU), Jukka Pontinen (Metla), Timo Siitonen (Metla), Esko Oksa (Metla), Jari Hietanen (Metla), Sami Lamminen (Metla), Marko Erkkilä (Techila oy), Freed Ahmad (UTU) and Derek Ross (UTU).

REFERENCES

Ala-Ilomäki, J. 2013. Spiked shear vane - a new tool for measuring peatland top layer strength Piikkisiipikaira - uusi väline turvemaan pintakerroksen lujuuden mittaukseen. Suo - Mires and Peat 64 (2-3), 113–118.

Beven, K. J. & Kirkby, M. J. 1979. A physically-based variable contributing area model of basin hydrology. Hydrology Science Bulletin 24 (1), 43–69.

Blaschke, T. 2010. Object based remote sensing, in ISPS Journal of Photogrammetry and Remote Sensing 65 (1), 2–16.

Blum Avrim, L. & Langley, P. 1997. Selection of relevant features and examples in machine learning, in Artificial Intelligence 97, 254–271. [Electronic resource]. Available at: http://yaroslavvb.com/papers/blum-selection.pdf . Last accessed 25 February 2015.

Böhner, J. & Antonic, O. 2009. Land-Surface Parameters Specific to Topo-Climatology. In Hengl, T.& Reuter, H. I. (Eds), Geomorphometry: concepts, software, applications, 195–226. Elsevier Science.

Burnash, R. J. C. 1995. The NWS river forecast system–catchment modeling. In: Singh, V.P. (Ed.), Computer Models of Watershed Hydrology. Water Resources Publications, Littleton, Colorado, 311–366.

Conrad, O. 2001–2010. Saga code implementation, 2001-2011. [Electronic resource]. Available at: http://www.saga-gis.org/en/index.html. Last accessed 25 February 2015.

Engheta, N. & Elachi, C. 1982. "Radar Scattering from a Diffuse Vegetation Layer over a Smooth Surface", Geoscience and Remote Sensing, IEEE Transactions on, vol. GE-20 (2), 212–216.

Grasty, R. L. 1997. Radon emanation and soil moisture effects on airborne gamma-ray measurements. Geophysics 62 (5), 1379–1385.

Hänninen, P. 1997. Dielectric coefficient surveying for overburden classification. Geologial survey of Finland, Bulletin 396. 72 p.

Hänninen, P., Lintinen, P., Lojander, S. & Sutinen, R. 2000. Suomen maaperän vedenjohtavuus. Vesitalous 41 (6), 16–19.

Hautaniemi, H., Kurimo, M., Multala, J. H. L. & Vironmäki, J. 2005. The three in one aerogeophysical concept of GTK in 2004. Geological Survey of Finland, Special Paper 39, 21–74.

Hyvönen, E., Pänttäjä, M., Sutinen, M.-L. & Sutinen, R. 2003. Assessing site suitability for Scots pine using airborne and terrestrial gamma-ray measurements in Finnish Lapland. Can. J. For. Res./Rev. Can. Rech. For. 33 (5), 796–806.

Kogan, R. M., Nazarov, I. M. & Fridman, Sh. D. 1971. Gamma spectrometry of natural environments and formations, Rep. 5778, Israel Program for Transl., Jerusalem. 377 p.

Kohavi, R. & John, G. H. 1997. Wrappers for feature subset selection, in Artificial Intelligence 97, 273–324. [Electronic resource]. Available at: http://ai.stanford.edu/~ronnyk/wrappersPrint.pdf. Last accessed 25 February 2015.

Koren, V. I., Smith, M., Wang, D. & Zhang, Z. 2000. Use

of SoilProperty Data in the Derivation of Conceptual Rainfall-Runoff Model Parameters Proceedings of the 15th Conference onHydrology. AMS, Long Beach, CA, 103–106.

McDonald, H. C., Waite, W. P. & Demarkce, J. S. 1980. American Society of Photogrammetry Annual Technical Meeting.

Middleton, M. 2014. Hyperspectral close-range and remote sensing of soils and related plant associations. Spectroscopic applications in the boreal environment, PhD Dissertation, Geological Survey of Finland. 68p. [Electronic resource]. Available at: http://tupa.gtk.fi/julkaisu/erikoisjulkaisu/ej_088.pdf . Last accessed 25 February 2015.

Miller, H. G. & Singh, V. 1994. Potential field tilt - A new concept for location of potential field sources. Journal of Applied Geophysics 32, 213–217.

Muro, T. & O'Brien, J. 2004. Terramechanics. Land Locomotion Mechanics. A. A. Balkema Publishers. ISBN 90 5809 572 X. 277 p.

Murphy, P. N. C. & Ogilvie, J. & Arp, P. 2009. Topographic modelling of soil moisture conditions: a comparison and verification of two models, Eur. J. Soil Sci. 60, 94–109.

Murphy, K. P. 2012. Machine Learning, A probabilistic perspective. The MIT Press.

Nevalainen, P. 2014a. Sensor Network Optimization 2014, UTU MSC thesis.

Nevalainen, P., Pohjankukka, J. et al. 2014b, Open Natural Resource Data in Forecasting the Harvester Mobility, In Federated Computer Science Event (YTP2014).

Oke, T. R. 1987. Boundary Layer Climates. 435 p. London, Taylor & Francis.

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andreassian, V., Anctil, F. & Loumagne, C. 2004. Which potential evapotranspiration input for a lumpedrainfall–runoff model? Part 2.Towards a simple and efficient potential evapotranspiration model for rainfall–runoff modelling. Journal of Hydrology 303, 290–306.

Päivänen, J. & Hånell, B. 2012. Peatland Ecology and Forestry – a Sound Approach. University of Helsinki Department of Forest Sciences Publications 3, 1–267. ISBN 978-952-10-4531-8.

Pennanen, O. & Mäkelä, O. 2003. Raakapuukuljetusten kelirikkohaittojen vähentäminen [Ways to reduce wood transport restrictions caused by the bad state of roads]. Metsätehon raportti. 153, 1–42. (In Finnish)

Pohja, T. 2000. Topi-proto penetrometri. Ohjeita käyttäjälle. T:mi Toivo Pohja. Käyttöohjekirja (Penetrometer instruction manual). In Finnish. 8 p.

Pohjankukka, J. 2013. Turun kauppatorin ilmanlaadun ennustaminen aikasarja-analyysilla, UTU MSc thesis.

Pohjankukka J. & Nevalainen P. 2014. Predicting water permeability of the soil based on open data. In: Iliadis, L., Maglogiannis, I. & Papadopoulos, H. (eds) Proceedings of the 10th International Conference on Artificial Intelligence Applications and Innovations (AIAI 2014), Volume 436 of IFIP Advances in Information and Communication Technology, Springer, 436–446.

Pohjankukka, J., Nevalainen, P. , Pahikkala, T., Hanninen, P., Hyvönen, E., Sutinen, R. & Heikkonen, J. 2014. Arctic soil hydraulic conductivity and soil type recognition based on aerial gamma-ray spectroscopy and topographical data. In: Proceedings of the 22nd International Conference on Pattern Recognition (ICPR2014).

Praks, J. 2012. Radar polarimetry and interferometry for remote sensing of boreal forest, Aalto Univ. Doctoral dissertations 153/2012. [Electronic resource]. Available at: http://lib.tkk.fi/Diss/2012/isbn9789526048734/isbn9789526048734.pdf . Last accessed 25 February 2015.

Quinn, P. F., Beven, K. J. P., Chevallier, P. & Planchon, O. 1991. The prediction of hillslope flow paths for distributed hydrological modelling using digital terrain models. Hydrological Processes 5, 59–79.

Rankinen, K., Karvonen, T. & Butterfield, D. 2004. A simple model for predicting soil temperature in snow-covered and seasonally frozen soil: model description and testing.Hydrology and Earth System Sciences 8 (4), 706–716.

Rehr, J. J. 2012. "High-performance computing without commitment: SC2IT: A cloud computing interface that makes computational science available to non-specialists", E-science 2012, IEEE 8th International Conference on E-Science,2012 (1–6). [Electronic resource]. Available at: http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=6404441&tag=1&abstractAccess=no&userType=inst. Last accessed 25 February 2015.

Richards, J. A., Woodgate, P. W. & Skidmore, A. K. 1987. "An explanation of enhanced radar backscattering from flooded forests", International Journal of Remote Sensing 8 (7), 1093–1100.

Seibert, J. & McGlynn, B. 2007. A new triangular multiple flow direction algorithm for computing upslope areas from gridded digital elevation models. Water Resources Research 43, W04501.

Sirén, M., Ala-Ilomäki, J., Mäkinen, H. , Lamminen, S. & Mikkola, T. 2013. Harvesting damage caused by thinning of Norway spruce in unfrozen soil. International Journal of Forest Engineering 24 (1), 60–75.

Sutinen, R. 1992a. Glacial deposits, their electrical properties and surveying by image interpretation and ground penetrating radar. Geological survey of Finland, Bulletin 359. 123 p.

Sutinen, R. & Hänninen, P. 1992b. Radar profiling and dielectrical properties of glacial deposits in North Finland. Proc. VI IAEG, 1045–1051.

Sutinen, R., Middleton, M., Liwata, M., Piekkari, M. & Hyvönen, E. 2009. Sediment anisotropy coincides with moraine ridge trend in south-central Finnish Lapland. Boreas 38, 638–646.

Tamminen, P. 1991. Kangasmaan ravinnetunnusten ilmaiseminen ja viljavuuden alueellinen vaihtelu Etelä-Suomessa. Summary: Expression of soil nutrient status and regional variation in soil fertility of forested sites in southern Finland. Folia Forestalia 777. 40 p.

Tomppo, E., Haakana, M., Katila, M. & Peräsaari, J. 2008. Multi-source national forest inventory - Methods and applications. Managing Forest Ecosystems 18. Springer. ISBN 978-1-4020-8712-7. 374 p.

Topo Toolbox, 2011. [Electronic resource]. Available at: https://topotoolbox.wordpress.com/ . Last accessed 25 February 2015.

Townsend, P. A. 2002. "Relationships between forest structure and the detection of flood inundation in forested wetlands using C-band SAR", International Journal of Remote Sensing 23 (3), 443–460.

Väätäinen, K., Lamminen, S., Ala-Ilomäki, J., Sirén, M. & Asikainen, A. 2013. Kuljettajaa opastavat järjestelmät koneellisessa puunkorjuussa - kooste hankkeen avaintuloksista. Metlan työraportteja / Working Papers of the Finnish Forest Research Institute 279. 24 p. (In Finnish with English summary).

Valjus, T., Säävuori, H. & Leväniemi, H. 2011. Pehmeikköjen paksuuskarttojen tuottaminen. Geological Survey

of Finland, archive report 11/2011, 27 p. Available at: http://tupa.gtk.fi/raportti/arkisto/12_2011.pdf. Last accessed 25 February 2015.

Virtanen, K., Hänninen, P., Kallinen, R.-L., Vartiainen, S., Herranen, T. & Jokisaari, R. 2003. Suomen turvevarat 2000. Geological survey of Finland, Report of investigation 156. 101 p. + 7 app.

Vukovic, M. & Soro, A. 1992. Determination of hydraulic conductivity of porous media from grain size composition. Water Resources Publications, Littleton, Colorado, USA. 83 p.

Whitebox Geospatial Analysis Tool. [Electronic resource]. Available at: http://www.uoguelph.ca/~hydrogeo/Whitebox/download.shtml. Last accessed 25 February 2015.

Wilson, J. P. & Gallant, J. C. (eds) 2000. Terrain Analysis - Principles and Applications. 520 p. New York, John Wiley & Sons, Inc.

Zevenbergen, L. W. & Thorne, C. R. 1987. Quantitative analysis of land surface topography, Earth Surface Processes and Landforms 12, 47–56.

Zhan, C. 1993. A Hybrid Line Thinning Approach, Minneapolis, Proceedings Auto-Carto 11, 396-405

APPENDICES

A1 THE FEATURES

First is the list of sites. Special field measured feature is mentioned per each site.

Table A1. The cases by site and the target feature.

id	location	subject	field acquired features
Pa	Parkano	hydraulic conductivity, top soil class	
Po1	Pomokaira	hydraulic conductivity exponent	soil samples for grain size analysis
Po2	Pomokaira	bearing capacity	ground radar response
Po3	Kemijärvi	aggregate deposits	geomorphological category
Pi1	Pieksämäki	rut depth	rut depth
Pi2	Pieksämäki	penetration resistance	penetration resistance
Pi3	Pieksämäki	drainage water network, location and condition	
Kol	Kolari	flood detection	

The following table establishes the id numbers for main feature groups, all are regular raster data. The identifier refers to following sections, when the scope of data used at each test area is being presented. The shorthand forms of site names are explained in Table A1.

Table A2. Primary features used in each test case

Sec.	name	grid c.	Pa	Po1	Po2	Po3	Pi1	Pi2	Pi3	Kol
-	constant 1		x	x	x	x	x	x		
4.1.1	MS-NFI (Metla forest inventory data)	20m		x	x		x	x	x	
4.1.2	peatland mask (Metla)	20m		x	x		x	x	x	
1.3	topsoil, groundsoil (GTK)	50m	x	x	x	x	x	x		
1.4	airborne potassium gamma window (GTK)	50m	x	x	x		x	x		
4.1.4	weather attributes (FMI)	10km					x			
1.5	aerial EM (GTK)	50m		x						
6	LiDAR (GTK;NLS?)	5-6/m^2				x			x	
7	COSMO-SkyMed (FMI)	10-50m	x							x
8	topographic height (NLS)	2-10m	x	x	x		x	x	x	x

A 1.1 Metla National Forest Inventory features

These are 43 features. 41 features are divided to the main tree species (Spruce,pine,Birch) and to attributes e.g.,trunk volume, branch biomass, age. Three features are categorical:

Table A3. Categorical NFI features

feature	range of values	description
Site fertility class	1–8	Grove...summit forest (*)
Land Class	1–3	Forest land/Poorly productive forest land/ Unproductive forest land
Site Main Class	1–4	Mineral /Spruce mire/Pine Bog/Open bog

A 1.2

The peatland mask: The coverage is whole Finland, the raster constant is 20 m and value is binary (0: less than 30 cm peatland vegetation, 1: over 30 cm). It was compiled by the Finnish Forest Research Institute using the open geographic information data derived from NLS Topographic database (NLS 2014). The NLS Topographic database is a dataset depicting the terrain and covering the whole of Finland. The positional accuracy of the Topographic database corresponds to that of scales 1:5 000 - 1:10 000 (NLS 2014).

The peatland mask consists of four different NLS Topographic database elements depicting different type of peatlands. These elements were first combined and then rasterized to 20 m grid using ArcMap software (ESRI 2014). The definitions for peatlands in the NLS Topographic database are:

1. area is mostly covered by peatland vegetation and
2. a minimum of 0.3 m peat thickness (NLS 2013: 44). A minimum criteria for area is 1000 m2.

In Upper Lapland thinner peatland can also be included to peatlands if the area is covered by peatland vegetation.

Links:

- NLS 2014: http://www.maanmittauslaitos.fi/en/digituotteet/topographic-database
- NLS 2013: http://www.maanmittauslaitos.fi/sites/default/files/Maastotietokohteet_2013.pdf (In Finnish)
- ESRI 2014: http://resources.arcgis.com/en/help/main/10.2/

A 1.3 Topsoil and groundsoil (GTK)

Both have c. 16 categories. Following is the list of categories and the relation between soil category and the corresponding hydraulic conductivity exponent *x*.

Table A.4. Soil types and hydraulic conductivity exponent. x_1 is the maximum hydraulic conductivity and x_2 is the minimum hydraulic conductivity (adopted from Hänninen et al. 2000)

id	Selite	x^1	x^2
1	Bedrock	10	12
2	Sandy till	3	6
3	Silty till	6	9
4	Gravel	2	5
5	Sand	4	7
6	Fine sand	5	7
7	Coarse silt	6	8
8	Fine silt	7	9
9	Clay	8	11
10	Gyttja	10	12
11	Carex peat	9	11
12	Sphagnum peat	8	10
13	Peat production area	3	12
14	Filling material	NaN	NaN
15	Water	12	12
16	Muddy silt	8	12
17	Muddy clay	8	12
0	NoData!	NaN	NaN

A 1.4 Airborne gamma

Airborne gamma was provided by GTK. The raster data is based on gamma-ray flux from potassium, which is the decay process of the naturally occurring chemical element potassium (K). This data indicates many significant characteristics of the soil, including the tendency to stay moist after precipitation and tendency to frost heaving. Also the soil type, especially density, porosity, grain size and humidity of the soil have an effect to gamma-ray radiation.

A 1.5 Aerial EM

Aerial electromagnetic data was provided by GTK. Primary AEM components, in-phase and quadrature, were transformed to apparent resistivity values by using a half-space model (Hautaniemi et.al. 2005). The apparent resistivity gives information on different kind of soil conductors. The apparent resistivity is governed by grain size distribution, water and electronic conductors content of soil and cumulative weathering.

A 1.6 COSMO-SkyMed

The COSMO-SkyMed constellation consists of four individual satellites equipped with X-band Synthetic Aperture radar (SAR) sensors. It is ideal for mapping natural hazards and damage, because it enables daily imaging of any region on the earth. COSMO-SkyMed offers several imaging modes, where the spatial resolution ranges between 0.5 m and 100 km. For higher spatial resolution, the image size is smaller. The satellite data is received in FMI's satellite reception center located in Tähtelä, Sodankylä.

A 1.7 LiDAR and Digital Elevation Models data

We downloaded elevation data from National Land Survey of Finland's file service for open data. Point density of this data is at least 0.5 points/m2, which is equivalent to approximately 1.4 m distance between points. Height accuracy of the product is 0.3 meters. NLS distributes the raw point cloud (.las) of the data, but also a rasterized Digital Elevation Model (DEM). https://tiedostopalvelu.maanmittauslaitos.fi/tp/kartta?lang=en

Several geomorphometric variables were calculated from the NLS DEM in SAGA GIS environment. We calculated aspect, potential solar radiation (diffuse + direct insolation), plan curvature, profile curvature, flow accumulation and slope. Moreover we calculated a Topograhic Wetness Index based on flow accumulation and local slope .

It seems that best ones for Machine Learning purposes are: height variation from local average, flow area, steepness, sun radiation input and local curvature approximants. These can be computed either by specialised tools like GIS modules, or independent software likt Toptoolbox [9].

Moreover, we created customized DEM's from .las-files in other substudies. Extraction of small channel features required finer resolution than 2 metres, so we used 0.5 m resolution instead.

Table A5. Features derived from topographic height

Variable	Algorithm specification	Original reference	Saga Implementation
Aspect			Conrad (2001-2010)
Slope	9 parameter, 2nd order polynomial	Zevenbergen & Thorne (1987)	Conrad (2010-2010)
Profile curvature	9 parameter, 2nd order polynomial	Zevenbergen & Thorne (1987)	Conrad (2010-2010)
Plan curvature	9 parameter, 2nd order polynomial	Zevenbergen & Thorne (1987)	Conrad (2010-2010)
Flow accumulation (*a*)	Triangular MFD	Seibert & McGlynn (2007)	Conrad (2010-2010), Grabs (2010)
TWI	ln(*a*/tanb)	Beven & Kirkby (1979)	Conrad (2010-2010)
Diffuse insolation		Böhner & Antonic (2009), Oke (1987) Wilson & Gallant (2000)	Conrad (2010-2010)
Direct insolation		Böhner & Antonic (2009), Oke (1987), Wilson & Gallant (2000)	Conrad (2010-2010)
local height	difference between Poisson average and local value	ULJATH , no ref.	Matlab code, ULJATH, no ref.
convergence	discrete approximation of local curvature	Topo Toolbox (2011) and link: http://csdms.colorado.edu/w/images/Usersguide_1_intro.pdf	Matlab code, ULJATH, no ref.

A 1.8 Meteorological weather data

Weather data provided by the FMI consisted of various weather attributes such as: mean temperature of the day, mean rainfall of the day etc. from years 2011-2013. In the analyses we used the mean temperature and mean rainfall data. The data was measured from 12 sites with each site having a grid size of 10 km x 10 km.

A 1.9 MS-NFI products

MS-NFI Dataset consist of 43 different variables (+ index). Currently the most recent available dataset which covers the whole of Finland is from year 2011 and the resolution is 20 m. An excellent description about the methodology used to create this product is given by Tomppo et al. 2008.

Table A6. MS-NFI features

MS-NFI Variable	Unit
Biomass, spruce, living branches	10 kg/ha
Biomass, spruce, stem residual	10 kg/ha
Biomass, spruce, roots, d > 1 cm	10 kg/ha
Biomass, spruce, stump	10 kg/ha
Biomass, spruce, dead branches	10 kg/ha
Biomass, spruce, stem and bark	10 kg/ha
Biomass, spruce, foliage	10 kg/ha
Biomass, broad-leaved trees, living branches	10 kg/ha
Biomass, broad-leaved trees, stem residual	10 kg/ha
Biomass, broad-leaved trees, roots, d > 1 cm	10 kg/ha
Biomass, broad-leaved trees, stump	10 kg/ha
Biomass, broad-leaved trees, dead branches	10 kg/ha
Biomass, broad-leaved trees, stem and bark	10 kg/ha
Biomass, broad-leaved trees, foliage	10 kg/ha
Biomass, pine, living branches	10 kg/ha
Biomass, pine, stem residual	10 kg/ha
Biomass, pine, roots, d > 1 cm	10 kg/ha
Biomass, pine, stump	10 kg/ha
Biomass, pine, dead branches	10 kg/ha
Biomass, pine, stem and bark	10 kg/ha
Biomass, pine, foliage	10 kg/ha
Site main class	Class, 1-4
Site fertility class	Class, 1-8
Land class	Class, 1-3
Land class based on FAO FRA	Class, 1-4
Stand age	Year
Stand mean diameter of	cm
Stand mean height	dm
Canopy cover	%
Canopy cover of broad-leaved trees	%
Stand basal area	m^2/ha
Volume, birch	m^3/ha
Volume, birch pulpwood	m^3/ha
Volume, birch saw timber	m^3/ha
Volume, spruce	m^3/ha
Volume, spruce pulpwood	m^3/ha
Volume, spruce saw timber	m^3/ha
Volume, other broad-leaved trees	m^3/ha
Volume, other broad-leaved trees pulpwood	m^3/ha
Volume, other broad-leaved trees saw timber	m^3/ha
Volume, pine	m^3/ha
Volume, pine pulpwood	m^3/ha
Volume, pine saw timber	m^3/ha
Volume, the growing stock	m^3/ha

www.gtk.fi
info@gtk.fi

The final report of the project "New computational methods for efficient utilisation of public data" deals with the possibilities of publicly available spatial data in mapping and predicting geospatial phenomena. The aim was to develop practical applications merely based on open data from the databases of the Finnish Meteorological Institute, Geological Survey of Finland, Finnish Forest Research Institute and National Land Survey of Finland. The techniques of geographic information systems, remote sensing and machine learning were applied to forest terrain trafficability for forest harvest machinery, the subgrade bearing capacity of forest roads, hydrological modelling, mapping of mass-flow aggregate deposits for infrastructure construction, quick-response mapping of forest floods and mapping of drainage networks.

Partners in cooperation:

GTK

Luke
NATURAL RESOURCES INSTITUTE FINLAND

FINNISH METEOROLOGICAL INSTITUTE

UNIVERSITY OF TURKU

All GTK's publications online at hakku.gtk.fi

ISBN 978-952-217-332-4 (PDF)
ISSN 0781-4240